UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

TRANSCRIPTOME PROFILING OF SPINAL MUSCULAR ATROPHY MODELS USING RNA-SEQ

Francisco Maria de Aboim Borges Fialho de Brito

DISSERTAÇÃO
MESTRADO EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL
ESPECIALIDADE EM BIOINFORMÁTICA

2014

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

TRANSCRIPTOME PROFILING OF SPINAL MUSCULAR ATROPHY MODELS USING RNA-SEQ

Francisco Maria de Aboim Borges Fialho de Brito

DISSERTAÇÃO
MESTRADO EM BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL
ESPECIALIZAÇÃO EM BIOINFORMÁTICA

Tese orientada pela Prof. Doutora Margarida Gama-Carvalho
e co-orientada pela Doutora Andreia Amaral

2014

This thesis resulted in the publication of an article and a poster.

Article

- Amaral AJ, Brito FF, Chobanyan T, Yoshikawa S, Yokokura T, Van Vactor D and Gama-Carvalho M (2014). *Quality assessment and control of tissue specific RNA-seq libraries of Drosophila transgenic RNAi models*. Front. Genet. 5:43. doi: 10.3389/fgene.2014.00043

Poster

- Francisco Brito, Andreia J. Amaral, Takakazu Yokokura, David van Vactor and Margarida Gama-Carvalho (2013). *A roadmap for tissue specific RNA-seq analysis of Drosophila shRNA models.* Molecular Biology in Portugal and EMBL (and EMBL Alumni), 18th July 2013, Lisboa, Portugal.

Abstract

Spinal Muscular Atrophy (SMA) is a neurodegenerative disorder that represents the second most common cause of hereditary infant death. It is caused by the reduced expression of the ubiquitous protein SMN (Survival of Motor Neuron), which is known to have a central function in the assembly of ribonucleoprotein complexes involved in pre-mRNA splicing. More recently, this protein has been reported to be involved in trafficking of mRNA molecules along neuron axons. Although the SMA causing gene has been identified for over a decade, the exact mechanisms that lead to the specific death of motor neurons remain unclear. A long-standing hypothesis suggests that the disease emerges from motor-neuron specific changes in pre-mRNA splicing that affect key genes required for the survival of these cells. A possible approach to identify these genes is whole-transcriptome profiling. Nowadays, one of the most powerful tools for transcriptome profiling is next generation RNA sequencing (RNA-Seq), which can provide data with minimal biological variation between replicates, resulting in a precise comparison of different phenotypes. However this is not a bias-free technique and library preparation and sequencing problems can introduce several artifacts which need to be addressed.

Here we present an RNA-Seq study of disease models based on *D. melanogaster* and *H. sapiens* iPSC cultures developed to help unravel the pathways related to SMN down-regulation and SMA by identifying changes in gene expression and transcript isoform expression.

During the development of the analysis pipeline for *Drosophila*, several difficulties were encountered, emerging from the inherent complexity of the process of preparing tissue specific RNA samples requiring dissection and pooling of multiple larvae brains, and the presence of an shRNA expression vector. This resulted in intra-treatment variance that needed to be addressed and stabilized. Furthermore, we found that some widely used algorithms for discovery of novel transcript isoforms can perform poorly on *Drosophila*, requiring the selection of alternative approaches. Similarly, the human analysis pipeline showed a high amount of variance due to the limited number of individuals used to create iPSC libraries, introducing bias in the analysis.

Finally, we showed via comparison of differentially expressed ortholog genes that changes caused by SMN down-regulation affect several conserved genes across species, making *Drosophila* a favourable approach for modelling SMA.

Keywords: RNA-Seq, Spinal Muscular Atrophy, shRNA, *D. melanogaster*, *H. sapiens*

Resumo

A Atrofia Muscular Espinhal (SMA) é uma doença neuro-degenerativa que representa a segunda causa mais comum de morte infantil hereditária. É causada pela expressão reduzida da proteína ubíqua SMN (Survival of Motor neuron), que tem uma função central na assemblagem de complexos ribonucleoproteicos envolvidos no *splicing* do pre-mRNA. Recentemente, esta proteína também foi reportada como estando envolvida no tráfego de moléculas de mRNA ao longo dos axónios. Apesar do gene causador da SMA estar identificado há mais de uma década, os mecanismos que levam à morte dos neurónios motores continuam por descobrir. Uma das hipóteses sugere que a doença emerge devido a alterações de eventos de *splicing* no pre-mRNA em neurónios motores, afectando genes chave necessários para a sobrevivência destas células.

Uma abordagem possível para identificar estes genes é o estudo do perfil do transcriptoma, que permite identificar como alterações de expressão génica e padrões de splicing podem levar à activação de vias moleculares relacionadas com a doença. Hoje em dia, uma das ferramentas mais poderosas para estudar o transcriptoma é a sequenciação de RNA usando métodos de "Next Generation Sequencing", mais conhecida por RNA-seq, que produz dados com o mínimo de variação biologica entre replicados, resultando em comparações precisas entre fenótipos. Esta técnica produz vários milhões de pequenas sequências de nucleótidos chamadas reads que, quando alinhadas a um genoma de referência, permitem quantificar os níveis de expressão do transcriptoma, servindo como a base de comparação entre diferentes condições. Porém, esta técnica não é livre de problemas de enviesamento e a preparação das bibliotecas e erros durante a sequenciação podem introduzir vários artefactos que necessitam de ser tratados.

Nesta tese apresentamos um estudo de RNA-Seq de modelos de doença de *D. melanogaster* e culturas de iPSC de *H. sapiens* desenvolvidas com o objectivo de caracterizar as vias biológicas relacionadas com a sub-expressão do SMN e as causas da SMA através da identificação de alterações na expressão génica e isoformas dos transcritos.

D. melanogaster é um modelo de doença que apresenta um grande número de genes conservados quando comparado a humanos. Também tem um sistema nervoso complexo, essencial para um estudo eficaz de doenças neurodegenerativas a um nível molecular. Devido a estas características, modelos de doença humana podem ser facilmente gerados em drosófila através do uso de vectores de expressão de "small hairpin RNA" (shRNA) para alterar a expressão do gene causador da doença homologo num tecido especifico ou em todo o organismo.

Para estudar os efeitos da sub-expressão de SMN em D. melanogaster foram produzidas bibliotecas de RNA-seq do sistema nervoso central de moscas com três

fenótipos diferentes: um controlo, um modelo baseado em shRNA que provoca um knock-down médio da expressão de SMN em neurónios e um modelo severo de SMA. Durante a análise, encontrámos vários problemas que emergiram da complexidade inerente do processo de preparar amostras de RNA de um tecido especifico, que requerem a dissecção e *pooling* de múltiplos cérebros de larva e da presença de um vector de expressão de shRNA. Isto resultou em variância entre replicados que precisaram de ser tratados e estabilizados. De modo a identificar se as bibliotecas estariam contaminadas por tecido não-neuronal, criámos um método baseado comparação dos níveis normalizados de expressão de genes com expressão específica em tecidos neuronais e expressão específica em tecidos não neuronais. Este método conseguiu identificar um padrão específico de expressão génica neuronal, dando-nos uma ferramenta para remover bibliotecas com um elevado nível de contaminação de tecido não neuronal.

Uma nova ronda de sequenciação mostrou reduzidos níveis de contaminação por tecido não neuronal porém, ao fazer uma análise de expressão diferencial entre os níveis de expressão dos genes do controlo vs SMA médio, SMN não foi encontrado como diferencialmente expresso. Concluímos que tal estaria associado ao facto do sistema nervoso central ser um tecido complexo e algumas das suas componentes, como as células da glia, não são afectadas pelo shRNA devido a não possuírem elav, necessário para a expressão do shRNA. Isto levou-nos a produzir uma terceira ronda de sequenciação, contendo o modelo severo de SMA, onde os niveis de expressão de SMN estão reduzidos em todos os tecidos. Nesta ronda encontrámos alterações no conteúdo de GC que foram demonstradas como estando relacionadas com duplicações de reads feitas pelas rondas de PCR antes de se efectuar a sequenciação. Neste modelo, a análise dos dados encontrou SMN como diferencialmente expresso bem como, ao efectuar uma “gene set enrichment analysis” (GSEA), vários processos biológicos enriquecidos relacionados com SMA e a sub-expressão de SMN. Uma análise de expressão diferencial de isoformas revelou várias alterações atribuídas à falta de expressão de SMN, previamente observadas em outros estudos. Por último também observámos que alguns dos algoritmos muito usados para a descoberta de novas isoformas de transcritos têm uma performance pobre em drosófila, requerendo a utilização de abordagens alternativas.

O modelo usado nesta tese para o estudo de SMA em humanos é baseado em iPSC (induced Pluripotent Stem Cells). São culturas de células estaminais estáveis desenvolvidas a partir de células somáticas adultas através da expressão de quatro genes (*Oct4*, *Sox2*, *Klf4* e *c-myc*) através de vectores virais, plasmídeos ou mRNA sintetizado codificando estes factores de transcrição. Estas culturas podem ser induzidas a diferenciar-se em vários tipos de células, tornando-as numa ferramenta extremamente poderosa no estudo de doenças genéticas humanas. Vários estudos usaram com sucesso esta abordagem para o estudo de

doenças neurodegenerativas como Parkinson e esclerose lateral amiotrófica.

Para estudar os efeitos da SMA nos humanos, foram produzidas bibliotecas de RNA-seq de neurónios motores diferenciados de culturas de iPSC criadas a partir de fibroblastos de um individuo normal (controlo), o mesmo individuo normal com um shRNA que diminui a expressão de SMN e um paciente de SMA. A análise dos dados de RNA-seq, de um modo semelhante a D. melanogaster, mostrou grandes valores de variância entre condições devido ao número limitado de indivíduos usados para criar as bibliotecas de iPSC, enviesando a análise. Esta análise levou à conclusão de que o modelo de shRNA não diminui os niveis como esperado, não tendo sido encontrado o gene SMN1 como diferencialmente expresso quando comparado com o controlo bem como a análise entre este modelo e o paciente de SMA mostra resultados similares à análise entre o controlo e o paciente. A análise entre o controlo e o paciente revelaram uma grande variação biológica relacionada com erros de amostragem, tendo sido encontrado cerca de um terço do genoma humano como diferencialmente expresso. Apesar disso, foi possível encontrar o sinal de uma resposta de expressão génica à falta de expressão de SMN via comparação com estudos de SMA prévios, bem como alterações de expressão de isoformas relacionadas com regulação de expressão snRNPs.

Por último, apesar dos vários desafios encontrados no processamento dos dados do modelo de drosófila e o modelo humano, mostrámos através da comparação de genes ortologos diferencialmente expressos que as alterações causadas pela sub-expressão do SMN afectam vários genes conservados entre estas duas espécies, mostrando que o modelo de *Drosophila* é uma boa abordagem para modelar a SMA. Também propomos modificações a fazer em futuros estudos usando estes modelos de modo a diminuir os erros de preparação das bibliotecas e sequenciação, melhorando a subsequente análise dos dados.

Palavras-chave: RNA-Seq, Spinal Muscular Atrophy, shRNA, *D. melanogaster*, *H. sapiens*

Acknowledgements

Dr. Margarida Gama-Carvalho

Dr. Andreia J. Amaral

Dr. David Van Vactor

Dr. Takakazu Yokokura

Everyone else who collaborated in the making of this thesis.

Index

List of Figures

List of Tables

List of Abbreviations

Adj-p – Adjusted P-value

ALS – Amyotrophic lateral sclerosis

BH – Benjamini-Hochberg

BP – Biological Process

BWA – Burrows-Wheeler Aligner

BWT – Burrows-Wheeler Transform

cDNA – complementary DNA

CNS – Central Nervous System

CNV – Copy Number Variants

DEA – Differential Expression Analysis

DE – Differentially Expressed

FPKM – Fragments Per Kilobase of transcript per Million mapped reads

GSEA – Gene Set Enrichment Analysis

GO – Gene Ontology

iPSC – induced Pluripotent Stem Cells

KEGG - Kyoto Encyclopedia of Genes and Genomes

MF – Molecular Function

NMJ – Neuromuscular Junction

PCA – Principle Components Analysis

QS – Quality Score

RNAi – RNA interference

RNA-Seq – RNA-Sequencing

SBS – Sequencing by Synthesis

shRNA – short hairpin RNA

SMA – Spinal Muscular Atrophy

WT – Wild-type

1 – Introduction & Objectives

1.1 – Transcriptome profiling

Transcriptome profiling has become a major focus in biological research, showing how changes in gene expression and splicing patterns can lead to the activation of molecular pathways related to disease. Indeed, it has shed some light into the processes involved in the cause of neurodegenerative diseases, such as Alzheimer disease, Parkinson disease and amyotrophic lateral sclerosis (ALS)[1,2].

RNA-Sequencing, or RNA-Seq[3], is currently the most widely used technology to sequence the transcriptome. It is based on the high-throughput sequencing (also known as deep sequencing, next generation sequencing or NGS) of a cDNA library generated from steady-state RNA, producing millions of short nucleotide sequences (30-400 nucleotides in size) also known as reads. These reads, when mapped to a reference genome, can be used to quantify transcript abundances and serve as the basis for comparison between different phenotypes, providing information about gene expression, sequence variation in the transcriptome, allele specific expression levels, and exon usage, as well as allowing for the identification of novel splice junctions and promoters. Also, since RNA-Seq is not limited to the detection of transcripts corresponding to a previously known genomic sequence, it allows the assembly of transcriptomes of species for which a reference genome is not yet available. However, like all sequencing technologies, it presents a series of challenges and disadvantages. Because the preparation of RNA-seq libraries relies on reverse transcription and PCR amplification before sequencing, several types of biases have been reported as a consequence, including random hexamer priming bias[4], GC content bias[5] and depletion of 3' and 5' ends of the transcripts, which impacts read nucleotide content and read annotation and bias the quantification of gene expression[6].

Creating an RNA-Seq library follows a protocol with several steps (Figure 1), the first one being the enrichment of mature mRNAs (which have a 3' poly-adenylated tail, or poly-A) from the total extracted RNA or the depletion of ribosomal RNA (rRNA). The mature RNA is then randomly fragmented by hydrolysis or nebulization and reverse-transcribed into a cDNA via random hexamer or oligo-dT priming. Alternatively, fragmentation can be done after the creation of the cDNA library. RNA fragmentation results in an even read coverage over the transcript body with a decrease in coverage towards the transcript ends, whereas cDNA fragmentation results in lower read coverage over the transcript body with an increase of read coverage on the 3' and 5' ends. The cDNA library is then size selected for fragments suitable for sequencing by one of the various high-throughput sequencing technologies available (e.g: Illumina's Genome Analyser/HiSeq, Applied Biosystems' SOLiD, Roche's 454

– reviewed in [7]), which generates reads from one (single-end reads) or both (paired-end reads) ends of each of the selected fragments, producing up to hundreds of millions of reads.

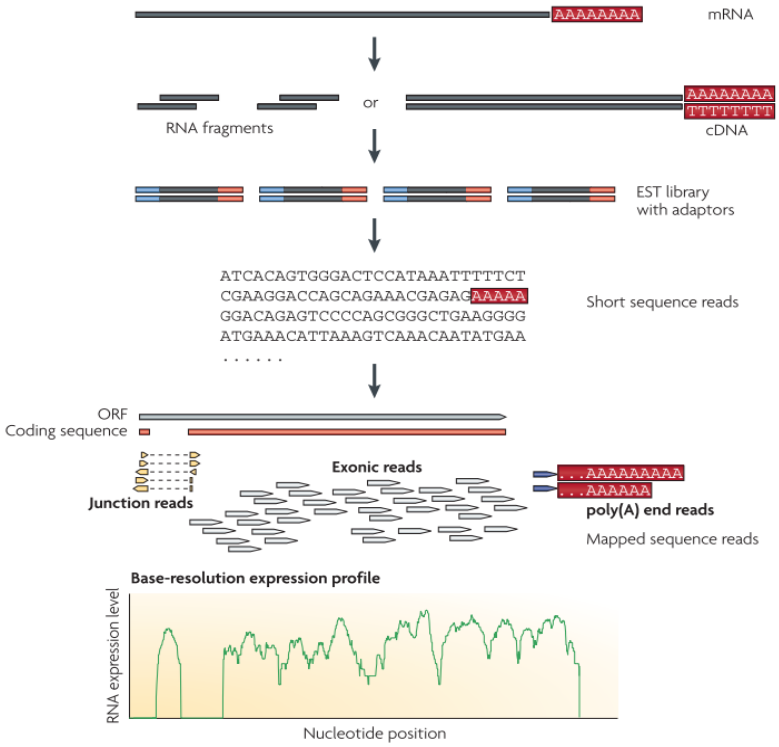


Figure 1 – A typical RNA-Seq experiment. Image and text adapted from Wang et al.[3]. "*Long RNAs are converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation. Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. The resulting sequence reads are aligned with the reference genome or transcriptome, and classified as three types: exonic reads, junction reads and poly(A) end-reads. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom.*"

1.2 – A Bioinformatical approach to RNA-Seq data

Depending on the scientific question of interest, the analysis of RNA-seq data will need different and sometimes very complex approaches, some of which tend to be very computationally heavy, such as aligning millions of reads to a reference genome or applying statistical models to test for differential expression of thousands of genes and respective exons, in order to answer questions such as which genes are being differentially expressed, or which isoforms are being enriched. This can be solved by using a bioinformatics approach, based on specialized algorithms for finding, matching and counting patterns in files with large volumes of data as fast and accurately as possible.

An RNA-seq data analysis requires several key steps: quality filtering of raw reads, mapping the reads to a reference genome, removing duplicate reads, quantifying genetic features, and finally comparing them between different experimental treatments (Figure 2). As mentioned before, RNA-seq data is biased by several errors introduced during either library preparation or due to the sequencing technique[4,5,6,8] which need to be filtered. Starting

out with the raw reads produced by one of the various existing NGS technologies, the first step is to filter out uncalled nucleotides, reads with a low Quality Score (QS), and reads with long homopolymers, which can bias the quantification of gene expression. Unidentified or uncalled bases are seen in reads as nucleotides tagged as N, identifying a position where it was not possible to accurately determine which nucleotide is being featured. Quality scores, or Phred scores[9], measure the probability of a base call being correct and are defined by the equation $QS = -10log10(e)$, where e is the estimated probability of the base call being wrong, meaning the higher the QS, the more accurate the nucleotide identification is. For example, if during the filtering process the QS is locked to values between 30 and 40, the probability of each base call being correct is at least 99.9% (QS=30) and at most 99.99% (QS=40). A higher or lower degree of stringency may be applied to reads based on the quality score, depending on the biological question. Lastly, homopolymers are a long repetition of the same nucleotide, usually covering 50% or more of the total read, which not only create noise during the base calling step, but are also ambiguous when aligning to a reference genome due their lack of a distinctive pattern, despite some of them being present in the actual transcriptome (ie: not caused by sequencing errors or library preparation errors).

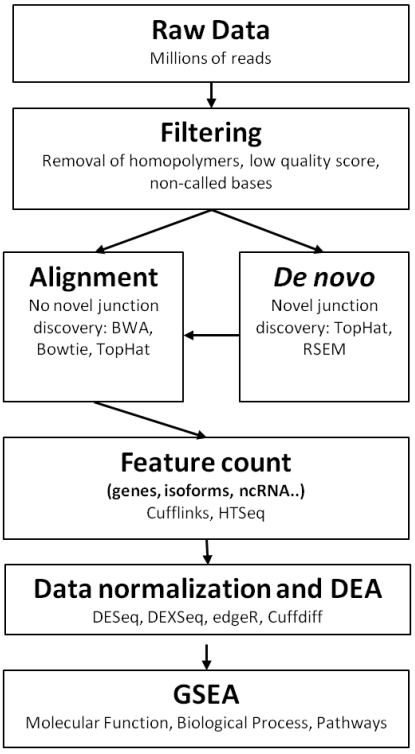


Figure 2 – A typical analysis pipeline for NGS data.

The second step consists of taking the filtered reads and determining their genomic coordinates by mapping them to a reference genome, using next generation aligners. There are two main types of next generation aligners: one uses algorithms based on hash tables

(e.g. BLAST, SOAP, MAQ) and the other uses algorithms based on suffix tries (e.g. Bowtie, BWA, SOAP2). Depending on the experimental design and/or biological question, these two algorithms have several different implementations to better suit the intended approach (reviewed in Li, 2010[10]). Suffix tries algorithms based on the Burrows-Wheeler transform (BWT) for suffix tries are the most used algorithms for RNA-seq data mapping. In comparison with traditional algorithms[11] (Figure 3a), BWT based aligners use a different data structure to store the seeds, called tries, which store all the suffixes of a sequence and compress them using an FM-index, a structure based on the BWT (Figure 3b). This results in multiple identical sequences that need to be aligned only once, increasing the alignment speed.

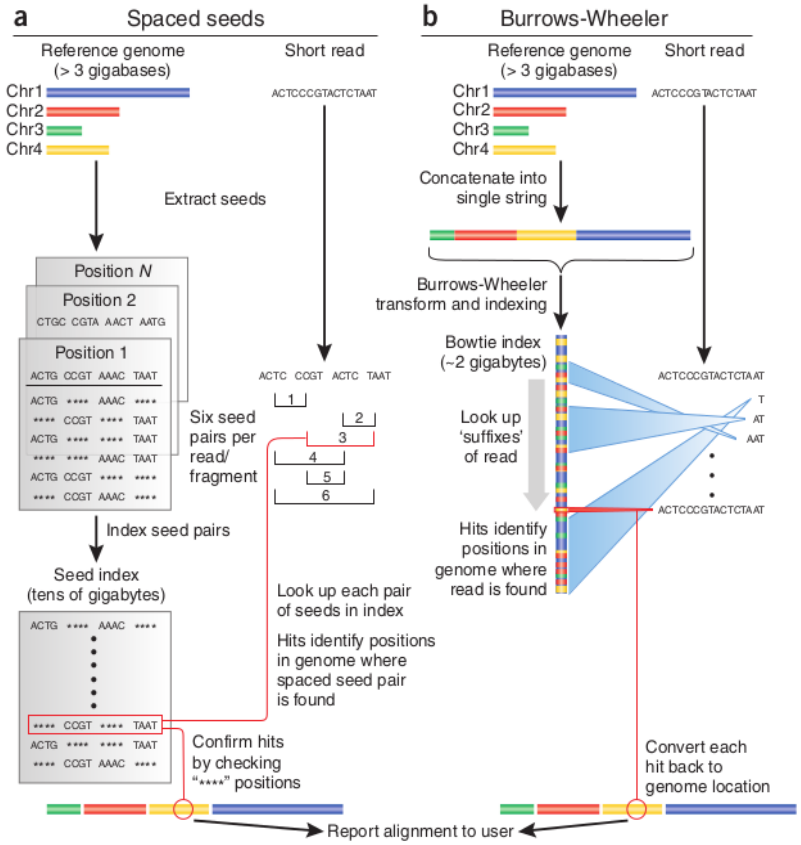


Figure 3 – Comparison between two different approaches for aligning short sequencing reads to a genome. Image taken from from Trapnell & Salzberg[12]. a) Spaced seed index approach b) Burrows-Wheeler Transform based approach.

Overall, software based on spaced seed indexing tends to be more sensitive but needs a large amount of memory and is less time efficient (30 fold slower)[13]. As an example, it takes 50GB+ of RAM to store a hash table of a human genome in memory, whereas software that uses algorithms based on the BWT can fit the same genome in under 2GB of memory. Having access to a server or computer grid capable of handling information on the scale needed by spaced seed software to store hash tables is restricted to most researchers and because of this, BWT based aligners are the preferred approach for a small server such

as the one being used in the analysis featured on this thesis. A well known example of this type of software is the Burrows Wheeler Aligner[14] (BWA), a widely used aligner which provides a cost-effective alignment of long reads (up to 1 Megabase) and allows for gapped alignment and nucleotide mismatches, but not reads that span splice junctions (Figure 4).

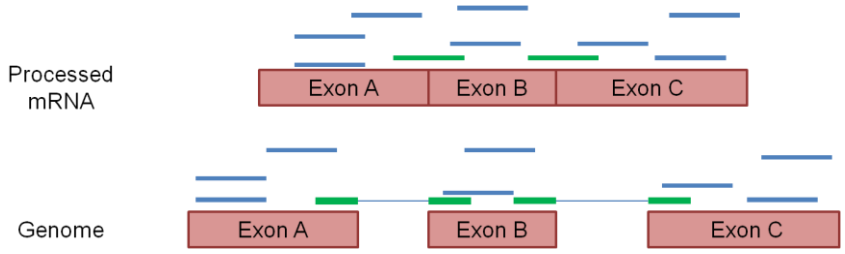


Figure 4 – Differences between aligning reads to the genome and the transcriptome. Represented in green are reads that can only be recognized by aligners that allow for detection of splice junctions. The reads' sequence contains part of two exons that have an intron separating them on the genome, which needs to be taken into account when aligning against the genome. The blue lines represent reads whose alignment spans only part of one exon and therefore aren't affected by intronic sequences when aligning to the genome.

Other software has been developed to address the limitations of the previous aligners, expanding their capabilities. One such example is TopHat[15], which takes RNA-Seq reads and compares them to the supplied reference genome in order to identify exon-exon splice junctions, based on the mapping provided by Bowtie[13], another BWT short read aligner which does not recognize reads overlapping splice junctions. With TopHat, reads are initially mapped by Bowtie to a reference genome, creating 2 files: one with mapped reads and another with unmapped reads. The unmapped reads are collected by TopHat and used to create a seed table index which allows the software to align them by taking into consideration that they can span splice junctions. This also allows for novel transcript discovery by aligning the raw data against the genome without giving any reference about intron/exon locations.

The third step comes in as an addition to the filtering step, as some sources of bias can only be detected after the reads have been mapped to the genome. As mentioned previously, during the RNA-Seq protocol, the RNA/cDNA is randomly fragmented, greatly lowering the probability of one of the generated fragments being exactly the same as another. If the same exact fragment is observed, it most probably indicates a duplication error, also known as PCR duplication, which results from overextending the number of PCR cycles needed during the amplification step, meaning the amplification process went beyond the point of saturation, resulting in duplicated fragments, or due to a high variance in fragment size, which results in an over-representation of smaller, more quickly amplified fragments. To detect them, specialized software such as SAMTools[16] and Picard[17] contain a feature that searches and filters out identical reads and/or reads that align in the exact same chromosomal coordinates.

Mapping the reads to the genome provides little information about the transcript abundance, and in order to extract information from the mapped reads, scripts have been

developed to count the number of reads mapping to the features contained in a given reference. This means you can count all reads which, for example, align against every known gene on a specific genome, compare their expression levels between two RNA-Seq libraries with different conditions (e.g.: control library vs disease library) and perform a differential expression analysis to better understand what qualifies as a condition-specific gene/isoform/other feature expression or simple biological variation. This step uses one of the numerous differential expression analysis (DEA) packages available (edgeR[18], DESeq[19], DEXSeq[20], Cuffdiff[21], SAMSeq[22], among others) to find which genes, isoforms or other features are being differentially expressed between conditions. Choosing the software will depend on the amount and type of data you have (reviewed in Soneson & Delorenzi[23]). Typically, these packages also contain some sort of data normalization process to reduce bias between different sized libraries and different gene lengths.

Finally, further data analysis can be performed with a gene set enrichment analysis (GSEA), which determines in which biological processes, molecular functions and pathways the DE genes/isoforms/etc are involved. With this information one can narrow down which pathways are being affected/triggered by the disease/treatment and derive some biological insight from the analysis.

1.3 – Disease models

Disease models play an important role in assessing which systems and pathways are affected by a certain disease. A commonly used model is *Drosophila melanogaster,* an organism that has been shown to have a great number of highly conserved genes when compared to humans[24]. It also has a complex nervous system, an essential characteristic for an effective study of neurodegenerative disorders at a molecular level[25,26]. Due to these features, human disease models can be easily generated in *Drosophila* through the use of small hairpin RNA (shRNA) expression vectors to target the *Drosophila* homologue of the disease causing gene in a tissue specific or organism wide manner, making it a very reliable and flexible approach to modelling disease. It can be used to create stable transgenic *Drosophila* mutant lines that can be used to address the impact of loss of function mutations on, for example, the central nervous system (CNS).

1.4 – Spinal Muscular Atrophy

Spinal Muscular Atrophy (SMA) is the second most common autosomal recessive disorder in humans, which presents itself by causing a degeneration of motor neurons[33], leading to an atrophy of the muscles and subsequently, respiratory insufficiency, paralysis and death. Its cause is known to be associated to the reduction of expression of the SMN (Survival of Motor Neuron) protein, due to a loss of function by the Survival of Motor Neuron

1 gene – *SMN1* – via deletions or, more rarely, missense mutations. This protein is known to have an important function in the assembling process of spliceosomal small nuclear ribonucleoproteins (snRNPs), but a link between its function and the degeneration of motor neurons has not yet been successfully established[34]. Nevertheless, it has been hypothesized that changes in pre-mRNA splicing mechanisms induced by a reduction of SMN expression affect other unknown genes, necessary for the survival of neuronal cells and neuromuscular junctions[35].

In humans, there are two SMN producing genes – *SMN1* and *SMN2* – which differ by a single nucleotide substitution in exon 7 (position 840, C to T), and are located in chromosome 5 in the telomeric and centromeric region, respectively. Even though this substitution is translationally silent, it prompts an alternative splicing event on SMN2, which skips exon 7 and causes about 85% of the proteins produced by SMN2's transcripts to be truncated, making it unable to compensate for the lack of SMN production in the event of a loss of function by *SMN1*. This implies that while the disease manifests itself due to *SMN1* loss of function, its severity is dependent on the number of existing *SMN2* gene copies, which determine how much functional, full-length SMN protein is being expressed.

On the other hand, *D. melanogaster* only has one SMN gene copy (Smn - chr 3L 16573498-16574647) and knocking down its expression results in lethality at the larval stage. Therefore, in order to mimic a human SMN1 loss of function, RNA interference (RNAi) is used to target the fly's *Smn* expression and reduce it to mimic SMN2 levels of expression. Previously existing studies on the cause of SMA have been based in null mutant models[36,37] where the SMN gene is disabled across all tissues, incurring in larvae lethality and thus not emulating the same conditions as a mild human SMA phenotype. By using a tissue specific driver to target only SMN production on the CNS, one can create specimens that more accurately profile the changes caused by this disease[38].

1.5 – Objectives

The work presented in this thesis aims to investigate how the genetic program of motor neurons is affected by the decrease of SMN expression, using RNA-Seq libraries from two disease models. The first one is an RNAi *D. Melanogaster* model that targets and lowers SMN expression in the central nervous system (CNS). The second model is based on iPS cell cultures differentiated from human SMA patients and healthy individuals. Furthermore it aims to assess the similarity between *D. melanogaster* and *H. sapiens* models, which should contribute with some insights regarding the usefulness of *D. melanogaster*'s models for the study of SMA disease.

2 – Materials & Methods

SMA *D. melanogaster* models, tissue preparation, RNA extraction and sequencing. Briefly, *Drosophila* transgenic RNAi models were developed by Dr. D. Van Vactor's group (HMS, US) using a binary system that uses the offspring obtained from a cross of two transgenic fly lines. The first line contained the neuron-specific *elav* promoter upstream of the yeast *GAL4* transcription factor coding sequence, while the second line had an integrated copy of the pWIZ vector with an intron-spliced hairpin transcript that produces a double stranded RNA for *Smn*, fused to the yeast upstream activator sequence (UAS) that is bound by GAL4 (Fig 5). Consequently, the offspring presents a neuronal-specific down-regulation of SMN, where *elav* was used as a tissue specific *GAL4* driver[39]. In essence, GAL4 is only expressed where *elav* is active (neuronal cells) and the RNAi is only transcribed when GAL4 is present. A parallel cross using a fly line containing an integrated copy of the empty pWIZ vector was used to generate a control wild-type (WT) strain for these studies.

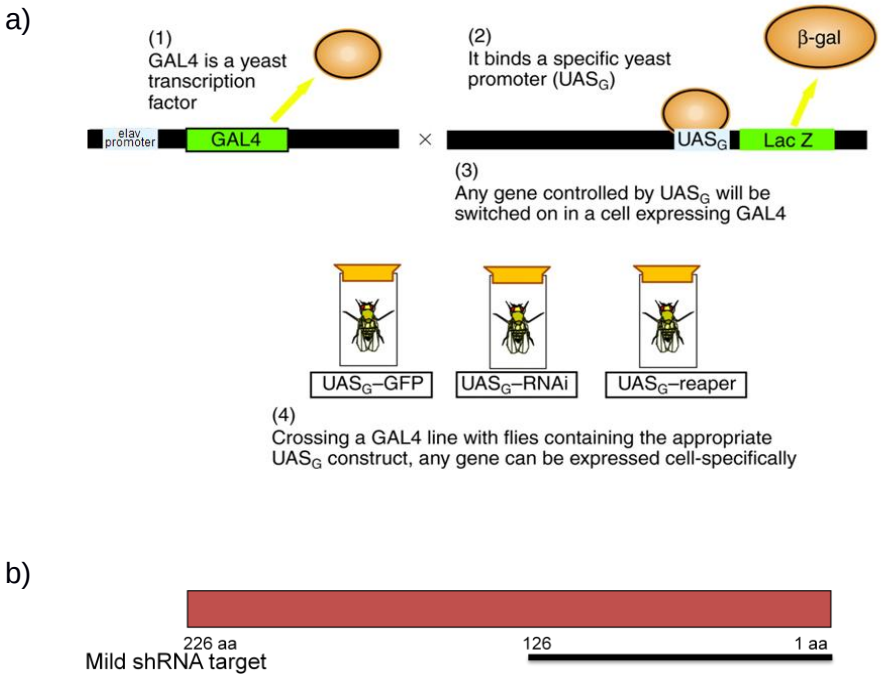


Figure 5 – The *elav-GAL4* model. a) The *elav-GAL4* construct (modifed image from Dow, Julian A T[40]) b) Representation of the *Smn* gene area targeted by C24.

Additionally, a null allele of *Smn* (Df(3L)*Smn*^{X7}) was used to generate a more severe knockdown of *Smn* levels, combining a 50% reduction of expression across all tissues with the neuronal specific *Smn* RNAi line (C24/X7)[38]. In total, four strains were used: w; P{w+mC=GAL4-elav.L}3 (Bloomington), w; P{UAS-PIWZ }15, w; P{UAS-SmnRNAi-C24}, and w; Df(3L)Smn^{X7}, P{UAS-SmnRNAi-C24}/TM6B, Dfd-YFP to create the wild-type and mutants, via crosses (Figure 6).

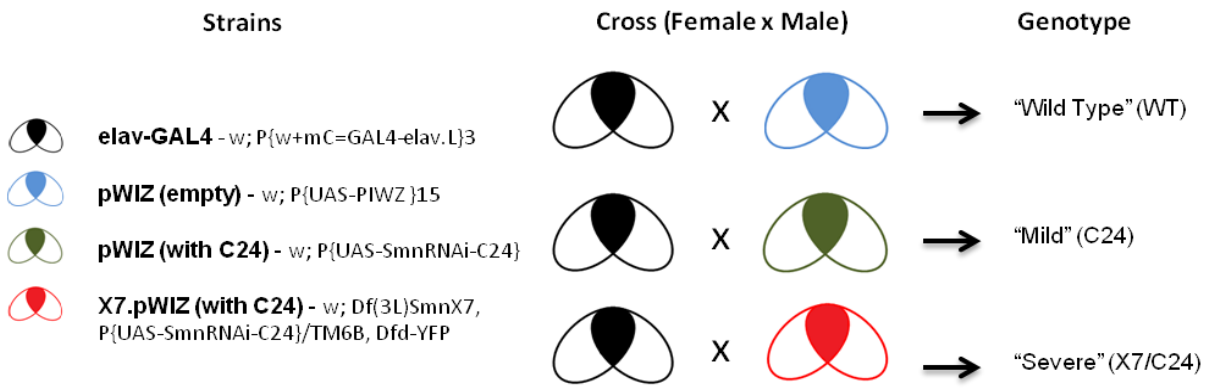


Figure 6 – *Drosophila* lines used in this study. Details the crosses made to obtain the studied genotypes (WT,C24 and X7/C24)

The central nervous system (CNS) of approximately 200 late third instar larvae was dissected in order to generate one biological replicate of the corresponding genotype. Dissected CNS samples were quickly frozen in TriPure Isolation Reagent (Roche Diagnostics GmbH, Mannheim, Germany) and pooled before performing total RNA extraction as described in Amaral et al. 2014[41]. In total, six biological replicates from WT flies, seven biological replicates from C24 flies and 4 biological replicates from X7/C24 flies were generated. mRNA-libraries were generated from an average 10μg of total RNA and prepared using the TruSeq RNA Sample preparation protocol (Illumina, USA). In summary, after two cycles of poly-A selection, RNA was fragmented to an average length of 300 bp and then converted into cDNA by random priming. The cDNA was then converted into a molecular library in order to generate paired-end RNA-seq libraries of 100 bp using the HiSeq2000 (Illumina, USA).

RNA-seq libraries of human iPSC-derived motor neuron cultures. Human motor neuron samples were produced by Dr Lee Rubin's group (Harvard Medical School) from retroviral generated iPSCs derived from type I SMA patient fibroblasts and control healthy fibroblasts as described in[42]. Furthermore, iPSCs derived from fibroblasts of the same healthy donor with an integrated shRNA vector targeting SMN1 were used in parallel to generate in vitro differentiated motor neurons (shSMN2). Total RNA was extracted and paired-end RNA-seq libraries were prepared as described. A total of 3 NS, 3 SMAiPS and 3 shSMN2 100bp RNA-seq libraries were sequenced using the HiSeq2000 (Illumina, USA).

Filtering, alignment and annotation. (See Figure 7 for the workflow) Raw data was analysed with FastQC to assess and visualize library's overall quality. Reads with homopolymers longer than 50 nt (>=50% of the read), non-called bases (tagged as N) and quality scores lower than 30 (QS < 30) were discarded using an in-house Perl script. The trimming of the first 10 nucleotides from each read (Appendix IV - Protocol 1) was made

using an in-house script in Python. Read alignment was made using the Burrows Wheeler Aligner v.0.6.1 (BWA) *aln* and *sampe* commands, allowing for only one mismatch and a distance of 200 nucleotides between read pairs, (Appendix IV - Protocol 2). Another aligner was also used TopHat[9] using two modes, one with the novel junction discovery feature on and another with novel junction discovery off. The genome assembly used for *D. melanogaster* was the BDGP5, and for *H. sapiens* was the GRCh37.71. SAMTools' "rmdup" feature[16] was used to remove potential PCR duplicates (Appendix IV - Protocol 2). Finally, gene counts were made with HTSeq's "*count*" feature in union mode[43]. In order to identify *GAL4* transcripts, all reads were mapped to *GAL4* gene (Gene ID 855828; accession NC_001148.4 - *Saccharomyces cerevisiae S288c*) using BWA[14] with the same previously described alignment parameters (Appendix IV – Protocol 3).

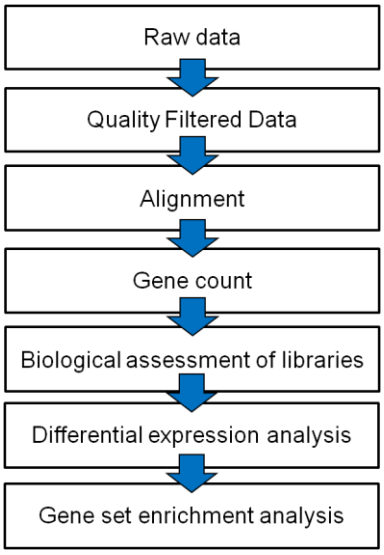


Figure 7 – Worfklow used for the analysis of the RNA-Seq libraries in both humans and flies.

Differential expression analysis and pathway enrichment. Data normalization was made using a size factor approach and tested for differential expression analysis by using bioconductor's DESeq[44] package, which encodes an algorithm based on the binomial negative model. DESeq was also used to for the biological assessment of libraries by correlating gene expression and variance values between libraries. Cuffdiff[21] was also used to test for differential expression of gene counts derived from the TopHat alignment. Out of the various available differential expression analysis softwares, DESeq uses one of the most statistically conservative methods and thus is less prone to find false positive results when performing differential expression tests[23]. Isoform DEA was made with DEXSeq[20], DESeq's equivalent for studying isoform expression. Clustering analysis was performed using the heatmap function from the ggplot package (default parameters) and correlation plots were generated using the lattice package in R environment[45]. Gene set enrichment analysis was

performed using "*GOstats*"[46], a bioconductor R package, in conjunction with other bioconductor packages, including "*org.Dm.eg.db*"[47], "org.Hs.eg.db"[48], "*KEGG.db*"[49], "*GO.db*"[50], "*biomaRt*"[51] and "*multtest*"[52].

Co-expression of ortholog genes between different models. Using BioMart[53] and the HGNC[54] database, a list of human/fly ortholog genes holding a HUGO ID was obtained. From this list, all genes that were previously identified as DE on both species were selected and used to generate a list of ortholog DE genes present in both species (Figure 8). Finally, a gene set enrichment analysis was performed using the GOstats R package[46] in order to understand which GO terms were being enriched by the DE ortholog genes. Note that only genes which were equally up-regulated or down-regulated in both species were used for the GSEA.

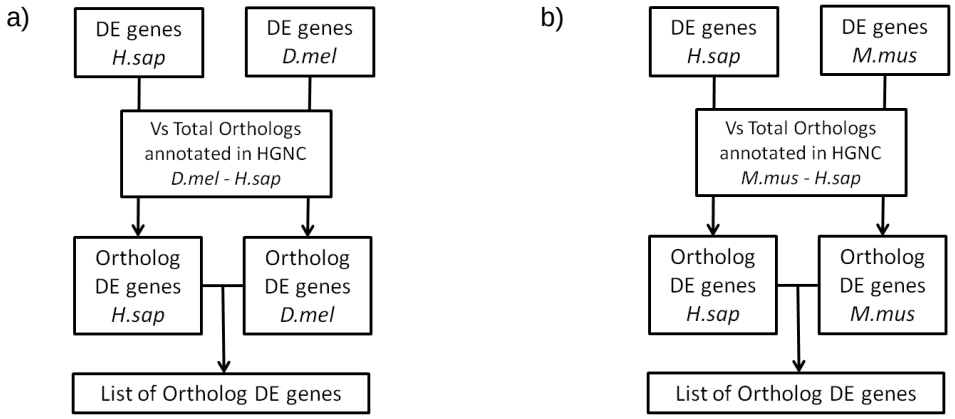


Figure 8 – Pipeline for the analysis of co-expression of a) *D. melanogaster* and *H. sapiens* ortholog genes and b) *M. musculus and H. sapiens* ortholog genes

GC content assessment between introns and exons. GC content for each library was quantified using an in-house python script which counts and separates reads according to the percentage of GC content. All read counts were normalized according to library size and the ratio between the GC content in aligned introns/genes across conditions was calculated.

3 – Results & Discussion

3.1 – *D. melanogaster* model

To study the effects of SMN's down-regulation on *Drosophila*'s CNS we analysed three distinct phenotypes. This analysis had the aim of uncovering which genes, exons and related pathways are affected by this down-regulation in this model and how it related to other studies using different *Drosophila* models, as well as its similarity with human SMA patients and it followed the pipeline described in methods, which includes data quality filtering of the generated datasets, alignment to a reference genome, differential expression analysis of genes and exons, and finally gene set enrichment analyses.

3.1.1 – Quality assessment of mRNA-seq libraries derived from the CNS of fly larvae

The datasets used in this study originated from three successive rounds of sequencing. The first sequencing batch is comprised of four biological replicates of CNS samples from a mild *Smn* knock-down mutant (C24) and three replicates of wild-type CNS (WT). The second batch contains three C24 and three WT replicates and the third batch has four biological replicates of the severe *Smn* knock-down mutant (X7/C24).

In the first batch/pilot run, we obtained an average of 100 million raw reads per library (Appendix I – *D. melanogaster*), 80% of which passed the initial quality filtering process. On average, 90% of the filtered reads were aligned to the genome using BWA, and ~75% of them were counted as part of a protein coding gene. After filtering, nearly 90% of the transcriptome was covered, with a sequencing depth of 500X and with an average of 11400 protein coding genes expressed, out of a total of 13872. In addition to the alignments made with BWA, we also used TopHat in this study, as it is one of the most widely used and well regarded softwares for alignment. Results show however, that in comparison with BWA and, despite TopHat having mapped an additional 5% of quality approved reads, the proportion of pairs that were accurately mapped was 20% less. Since TopHat has a lower coverage, this lead to the decision of using BWA as the preferred aligner to produce results for the downstream analysis.

We also observed that the percentage of read duplicates in C24 3 and WT 3 were much higher in comparison to the other replicates, indicating a problem in the library generation step, probably due to PCR saturation. As previously mentioned, PCR duplication is an issue that can lead to a skewed downstream analysis, since it introduces bias in gene expression by changing their expression levels due to the high number of read duplicates, which are being mapped as regular reads and counted as part of the gene expression levels.

To confirm if the libraries display the expected correlation with their assigned

phenotype, we performed a hierarchical clustering analysis and a correlation analysis, based on the whole-genome expression profiles. Results show that samples did not cluster according to phenotype (Figure 9), nor did the correlation values show consistency between libraries. For example, the WT1 dataset shows a higher correlation to the C24 libraries than to the WT 2 library. Given the technical complexity involved in isolating larvae CNS samples for RNA-seq profiling, we hypothesized that the presence of non-CNS tissue might be the major cause of the observed discrepancies, and that the Imaginal Discs (ID) could represent the largest source of contamination due to their proximity to the CNS at the studied larval stage.

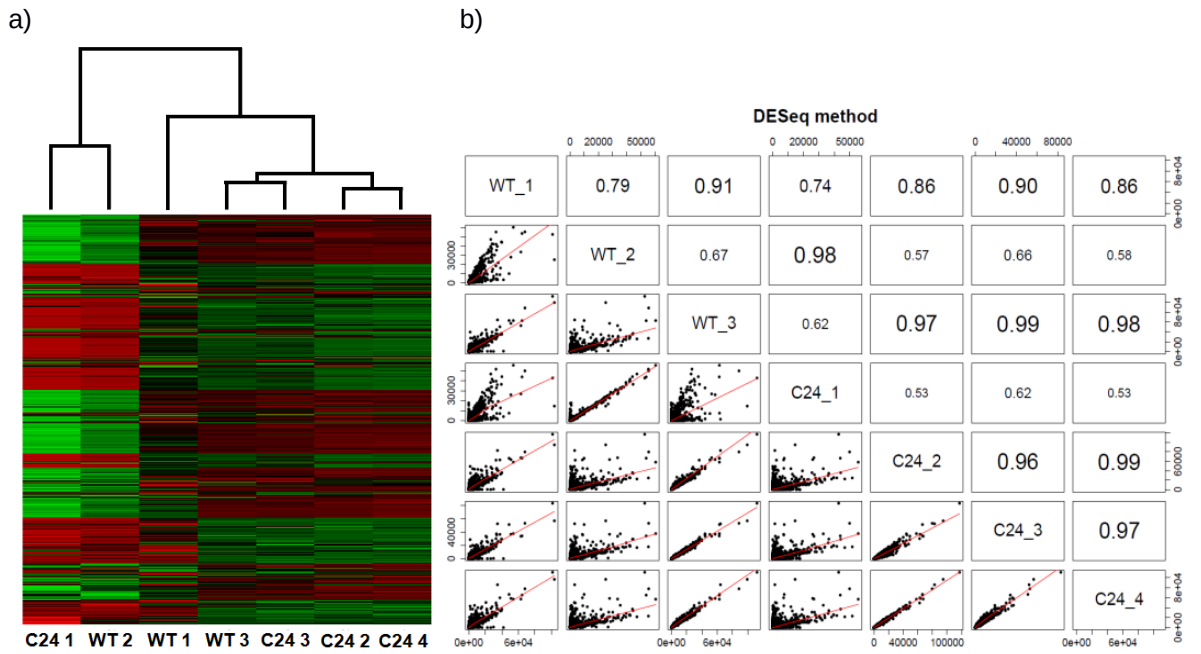


Figure 9 – Clustering and correlation analysis of the first sequencing batch. C24 1 to 4 represent the datasets generated for the mild *Smn* knockdown mutant, while WT 1 to 3 represent the control. a) Clustering between all samples from the first sequencing batch based on stabilized variance. The plot shows that the datasets are clearly not clustering according to phenotype; b) Correlation plot between all samples from the first sequencing batch, based on gene expression values normalized by DESeq (the higher the value, the higher the correlation is between libraries).

We further hypothesized that by selecting a set of key marker genes it could possible to identify mRNA-seq libraries that were displaying levels of gene expression closest to the expected for CNS derived samples. Using FlyBase's high-throughput expression data[55], we selected five genes: two tissue specific genes (*elav*, *repo*) and three ubiquitous genes (*Pen*, *Usp7* and the RNAi target gene, *Smn)*. *elav* has a neuronal specific expression while *repo* is reported has having a CNS glial cell specific expression. Although ubiquitous, *Pen* has a higher expression in the imaginal discs than on the CNS and *Usp7* is evenly expressed across most tissues. Additionally, the transgene driver *GAL4* was also selected to be included in this profile. *GAL4*'s expression is regulated by the presence of *elav* and is not present on wild type *D. melanogaster* (Table 2). *Smn*, *elav*, *repo*, *Pen* and *Usp7* expression

levels were obtained for all libraries by quantifying the paired-end reads aligned to the *Drosophila* genome with HTSeq and then normalized for gene length. *GAL4* expression levels were quantified by aligning the RNA-Seq libraries to a "one-gene genome" created from the *GAL4* gene fasta sequence and counting the uniquely aligned reads. A comparison of the expression levels of these six genes across all libraries from the first sequencing batch (Figure 10) confirmed the hypothesis that several libraries display a non-neuronal expression pattern (namely WT 1 and 2 and C24 1), giving a basis for excluding these from the dataset, as well as defining criteria for control of biological origin for further sequencing rounds.

Table 1 – Gene expression levels of the benchmarking genes described in Flybase's high-throughput expression database and their respective gene length for the central nervous system (CNS) and imaginal discs (ID). Note that *GAL4* is a yeast gene introduced for the purpose of the experiment and does not have any kind of gene expression in a regular fly. The values represent the modENCODE's expression level measure: very low expression (1-3), low (4-10), moderate (11-25), moderately high (26-50), high (51-100), very high (101-1000) and extremely high expression (>1000)

Expression levels	Pendulin	Smn	Usp7	elav	repo	GAL4
CNS	111	76	71	82	25	No expression
ID	498	132	58	6	3	No expression
Gene Length (bp)	3189	876	6200	10763	3408	2645

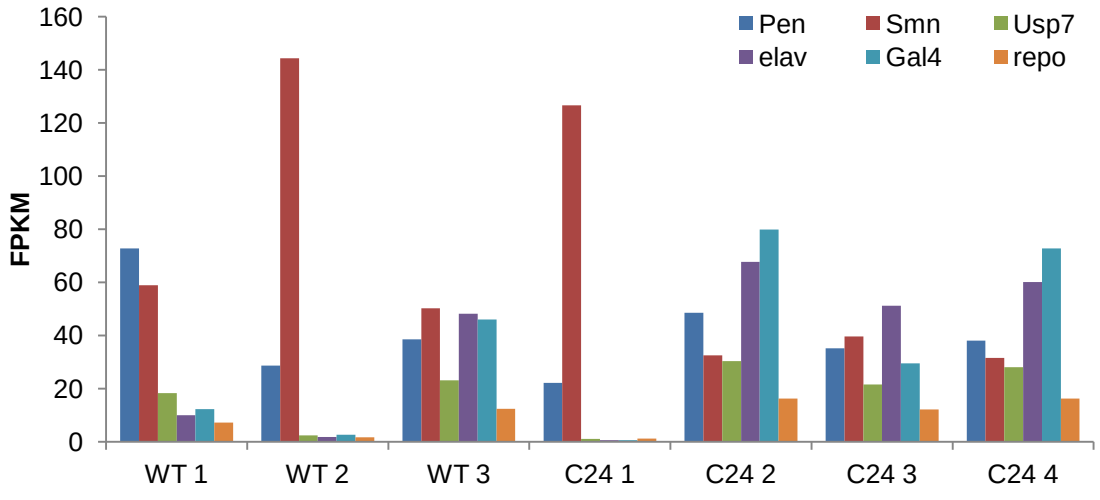


Figure 10 – Benchmarking genes' expression values for the first sequencing batch libraries, normalized for library size and gene length. WT 3 and C24 2-4 gene expression values correspond to neuronal tissue expression. All other samples (WT 1, WT 2, and C24 1) display almost no *elav* expression (neuron specific gene). WT 2 and C24 1 also display very high levels of SMN, expected from tissue derived from imaginal discs.

We also observed that GAL4 expression levels did not follow *elav*'s expression levels in samples C24 3 and WT 3, the two libraries that had been previously observed as having a

higher amount of PCR duplicates. In fact, when performing the clustering analysis with the samples that passed the tissue specific analysis, these two samples did not cluster as expected, showing a clear separation between libraries according to the amount of duplicates (Figure 11). These results supported the final decision that the pilot sequencing round did not provide data with sufficient quality for performing a differential expression analysis and that therefore a new sequencing round would have to be performed.

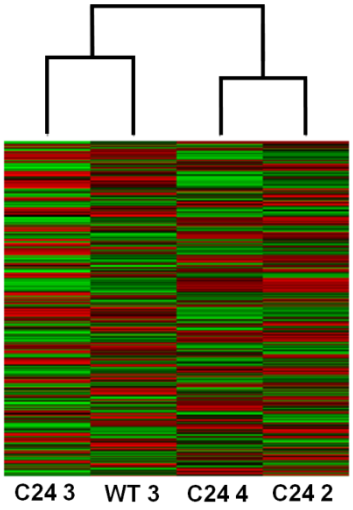


Figure 11 – Clustering analysis of the four libraries that passed the benchmarking gene assessment. Samples are being clustered according to the amount of PCR duplicates present on each dataset rather than by phenotype (~60% in C24 3 and WT 3, ~40% in C24 4 and C24 2).

3.1.2 – Characterization of the transcriptome of the CNS of neuronal *Smn* knockdown *Drosophila* lines modelling a mild SMA phenotype

A second batch of data was produced containing three new replicates of CNS samples from wild type and C24 larvae. In this batch, an average of 101,7 million raw reads were obtained per library (Appendix I – *D. melanogaster*), 84,5% of which passed the initial quality filtering process. On average, 91,4% of the filtered reads were aligned to the genome using BWA, and ~76,7% of them were counted as part of a protein coding gene. After filtering, nearly 90% of the transcriptome was covered, with a sequencing depth of 500X and with an average of 11824 protein coding genes expressed, out of a total of 13872. PCR duplication accounted for an average of 28% of the uniquely aligned reads, much less than the duplication rates found on the first sequencing batch.

The gene panel described in the previous section, used to evaluate library quality regarding their tissue of origin was also used to assess this sequencing batch. In this batch, all samples present the same kind of expression pattern previously found to be indicative of neuronal tissue, suggesting that these samples presented a much smaller degree of contamination from neighbouring tissues (Figure 12). However, the clustering analysis failed once more to classify these libraries according to their origin, with C24 6 and WT 5 being

clustered separately from the other libraries (Figure 13).

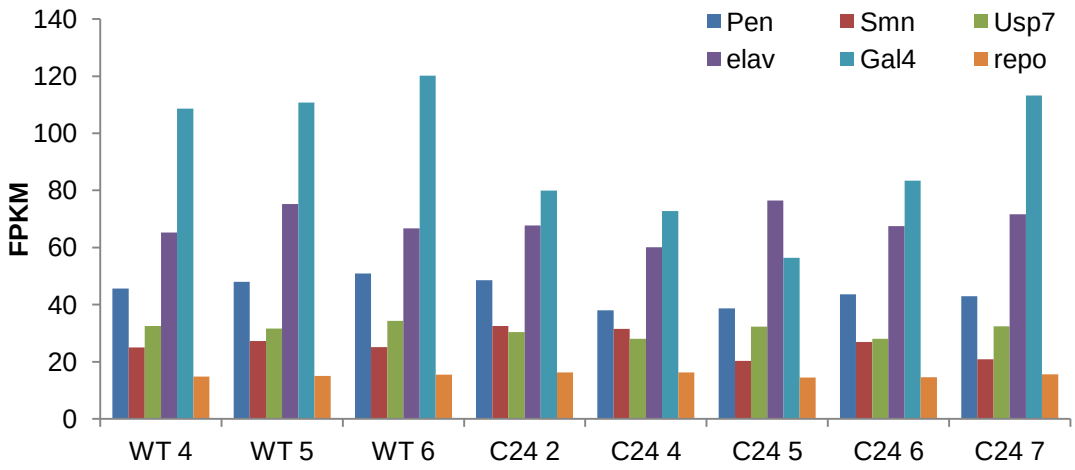


Figure 12 – Expression values of the benchmarking genes from the second sequencing batch and the first sequencing batch samples that passed the test.

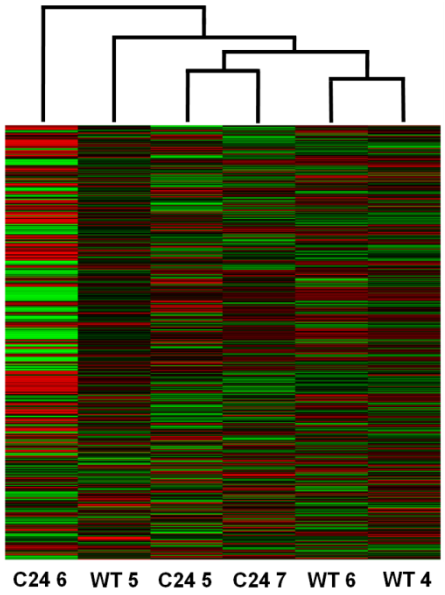


Figure 13 – Clustering analysis of the second sequencing batch. As seen, C24 6 and WT 5 are not clustering according to phenotype.

Following these observations, we decided to investigate if these samples were being affected by a slightly higher presence of other non-neuronal tissues when compared to the other libraries. As before, using the public repository of RNA-seq data in Flybase, we quantified the gene expression levels which are exclusive to the neuronal system, as well as genes that display a high expression in other tissues (imaginal discs, carcass, fat body, salivary glands and digestive system) but have low to no expression in the neuronal system. Results (see Appendix II) showed that WT 5 has higher expression levels on genes present in the imaginal discs, digestive system, carcass and salivary glands but not on the CNS. C24 6 on the other hand, had lower expression levels for CNS specific genes though it

maintained the same expression levels as the other samples on the non-CNS tissues. These altered levels of gene expression however, were not considered to be enough to skew the downstream analysis since both libraries passed the assessment for neuronal specific gene expression with almost the exact same expression pattern of the other replicates, suggesting both libraries should not be discarded. After the tissue-specific gene assessment, we performed a principal component analysis (PCA) with all the approved libraries from both sequencing batches (Figure 14). Again, results showed C24 6 and WT 5 as possible outliers, though a clear separation between WT and C24 can be observed.

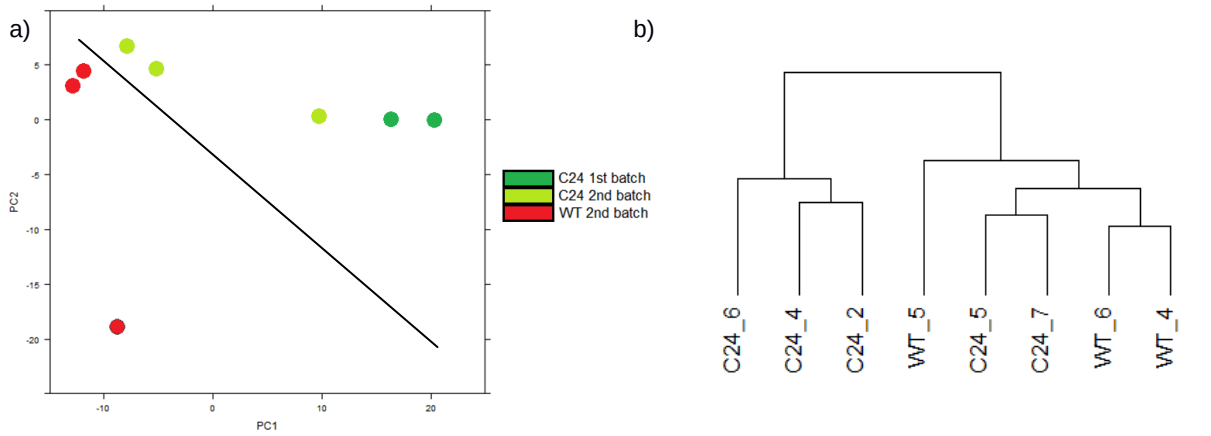


Figure 14 – Transcriptome profiling of the second sequencing batch. a) Principle Components Analysis (PCA) plot and b) Hierarchical clustering of the second sequencing batch libraries and the viable C24 replicates from the first sequencing batch.

Considering the apparent differences between the libraries from the second sequencing batch, we decided to perform two DE analysis, one using all six libraries (3x3 analysis), and another using only the two replicates for each condition that displayed good results in both the clustering analysis and the tissue-specificity assessment (2x2 analysis). Results show a relatively small number of significantly DE genes (adj-p > 0,05) in both approaches: 78 genes in the 3x3 analysis, 209 in the 2x2 analysis, 61 of which are common. Unexpectedly, neither analysis identified the shRNA target, SMN, as differentially expressed showing only a slight change in gene expression levels (Table 2).

Table 2 – SMN's expression fold change between WT and C24 libraries. Using three replicates for each condition (3x3) and removing the two replicates that did not clustering according to phenotype (2x2).

Approach	SMN fold change (log2)	Significant? (adj-p < 0,05)
3x3	-0,13	no
2x2	-0,19	no

Two possible reasons might explain this outcome. The first one is that SMN's expression levels in the C24 libraries were overestimated due to a bias caused by the assignment of

shRNA derived reads to the SMN gene. As previously explained, the shRNA knockdown of SMN is mediated by the expression of an antisense transcript corresponding to approximately half of the endogenous transcript (see Figure 5b). Therefore, we decided to investigate if a bias in the SMN read coverage towards the region complementary to the shRNA was observed. For this purpose, we used the results from the second sequencing batch to plot the distribution of all reads aligned to *Smn* across its chromosomal location in order to see if there was a significant difference in read distribution between the area targeted by the shRNA, C24, and the area that is not targeted by C24. If true, we also hypothesized that it would be possible to observe a difference in gene expression between WT and C24 in the area not targeted by the shRNA. Results indicate this is not the case (Figure 15a), suggesting that shRNA expression did not interfere with the target gene's quantification. This quantification of SMN's read coverage revealed that it is higher in the transcript body and decreasing towards the ends. This is what is expected in RNA-seq libraries generated from RNA fragmentation, as described in the methods section (Figure 15b). We therefore concluded that the shRNA used in this study was not biasing the estimation of SMN expression levels.

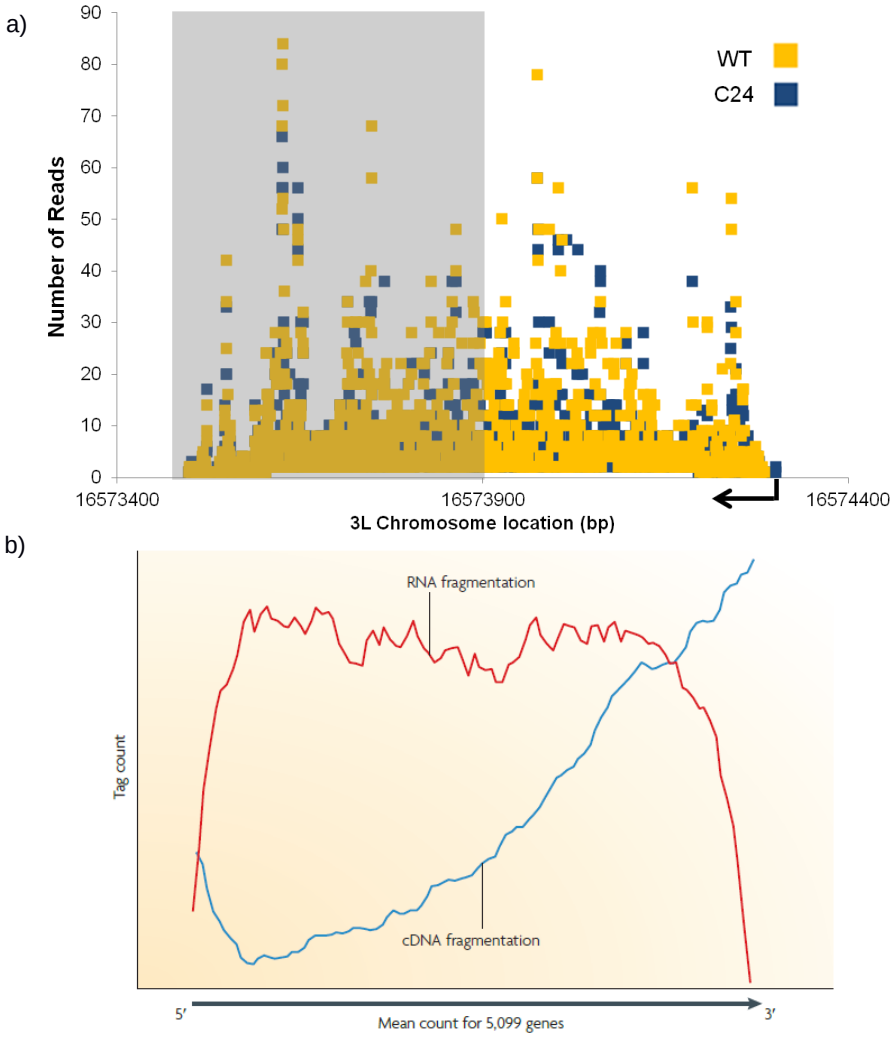


Figure 15 – SMN read distribution. a) Distribution of the reads aligned to the SMN gene in the second sequencing batch by their chromosomal location. The grey area indicates C24's target and the black arrow indicates *Smn*'s start site; b) Taken from Wang et al.[3] - Read distribution along a gene according to the type of fragmentation used in the library preparation.

The second possibility was based on a lack of a significant reduction of SMN expression in C24 libraries, probably due to the shRNA not having an impact on SMN expression as great as expected. To test this, we perfoemed a blind DE analysis by doing a DEA to random combinations of the data set: three 2x2 and three 3x3. Results show an average of 20 DE genes in the 3x3 analysis and an average of 14 DE genes in the 2x2 analysis with none of them overlapping either WTxC24 gene DEAs. These results suggested that the DE genes found on the WTxC24 DEAs are not a result of random biological variation across all samples but of the SMN-dependent variation found between the two conditions (WT and C24).

Having dismissed both possibilities, we put forth a third hypothesis based on the fact that the central nervous system is a complex tissue comprised of several cell-types, including glial cells. These cells do not express *elav,* which is necessary to activate the expression of the shRNA that targets *Smn* and down-regulates its expression levels. Since the C24 shRNA model has been shown to work in studies where its expression is ubiquitous[38], the presence of the glial cells in these libraries coupled with the tissue specific driver implemented in this fly model could explain why SMN levels are not being significantly affected, as the SMN expression stemming from the glial cells is not being targeted by the shRNA, masking the down-regulation caused on neuronal cells. At this point however, it was not possible for us to determine if the presence of glial cells had a significant impact in the expression of SMN without a phenotype where *Smn* is ubiquitously down-regulated.

3.1.3 - Characterization of the CNS transcriptome of *Drosophila* lines with neuronal *Smn* knockdown on a heterozygous null background

Since the previous results lead us to the hypothesis that the effect of the shRNA on SMN expression levels proves to be difficult to detect in a complex tissue such as the CNS, the creation of new libraries derived from a more severe SMA *D. melanogaster* model would be required to further understand the effects of SMN down-regulation on the CNS transcriptome. To that effect, a third sequencing batch was produced, including four replicates of the severe SMA *Drosophila* model (X7/C24). This model presents a null mutation of the *Smn* gene (X7) in heterozygosity, over which the neuronal specific C24 shRNA was introduced.

In this third batch we obtained an average of 100,7 million raw reads per library (Appendix I – *D. melanogaster*), 86% of which passed the initial quality filtering process. On average, 92,2% of the filtered reads were aligned to the genome using BWA, and 75,5% of

them were counted as part of a protein coding gene. After filtering, nearly 90% of the transcriptome was covered, with a sequencing depth of 500X and with an average of 11771 protein coding genes expressed, out of a total of 13872. PCR duplication accounted for an average of 30% of the uniquely aligned reads (similar to what we found in the second sequencing batch) except for the fourth replicate (named X7/C24 D), which was flagged for having a much higher percentage of duplicates (~50% of the mapped reads flagged as duplicates). We believe this may have been caused by a lower cDNA input used for X7/C24 D which lead us to remove this library from the downstream analysis, followed by an excessive number of PCR cycles which saturated the sample, producing an mRNA-seq library with low transcript complexity (see Appendix I – *D. melanogaster*). All the other samples passed the hierarchical clustering assessment and showed a clear separation between conditions when compared with the WT libraries from the second batch (Figure 16). Furthermore, the tissue specific analysis showed that all four samples displayed a gene expression pattern corresponding to CNS tissue (Figure 17).

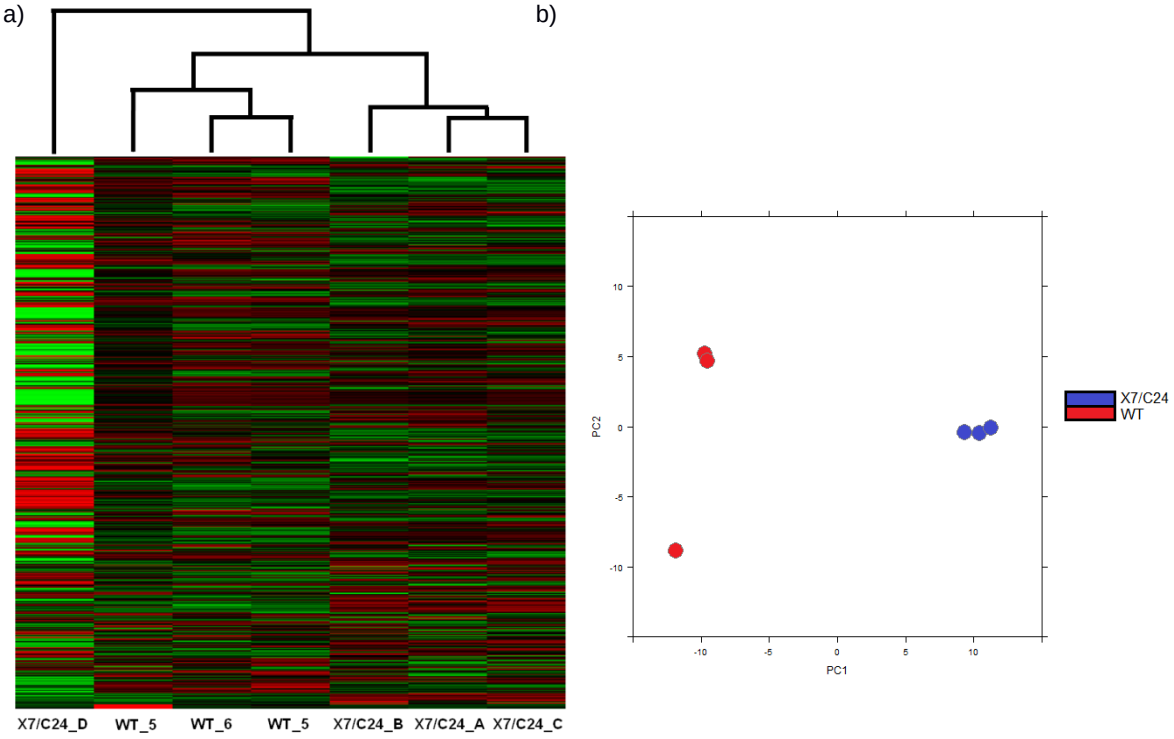


Figure 16 – Transcriptome profiling of the X7/C24 dataset (third sequencing batch). a) Hierarchical clustering between X7/C24 and WT (2nd batch); as seen, X7/C24 D is clearly not clustering according to phenotype. b) PCA plot comparing the WT libraries from the second batch and the X7/C24 libraries that passed the quality filtering step. A clear separation between both phenotypes can be observed.

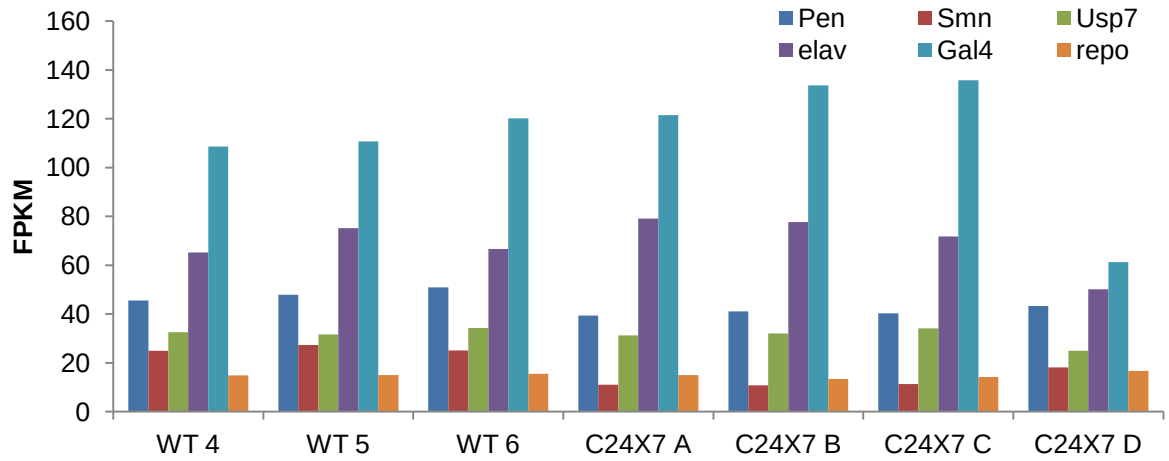


Figure 17 - Expression values of the benchmarking genes from the second sequencing batch and third sequencing batch samples, normalized for gene length and library size. All X7/C24 datasets present a neuronal tissue gene expression.

As seen below (Figure 18), we also observed unusual increase of GC content in mRNA-seq libraries derived from the severe SMA *Drosophila* model in reads containing between 19 and 40% GC content. Since introns have been shown to have a lower GC content than exons[56], we hypothesized the origin of this peak could be related with an increased intron retention caused by *Smn*'s down-regulation due to its function on snRNP assembly. Alternatively, it could be related with a contamination by another species, such as bacteria, which are known to have widely varied levels of GC content[57]. To test this, we analysed the alignment percentage in the 19-40% GC content area in all three conditions (WT, C24 and X7/C24) using the datasets from the second and third sequencing batch, and we searched for reads from this interval that aligned to all known *Drosophila* introns in order to assess the ratio between introns and reads mapped to genes.

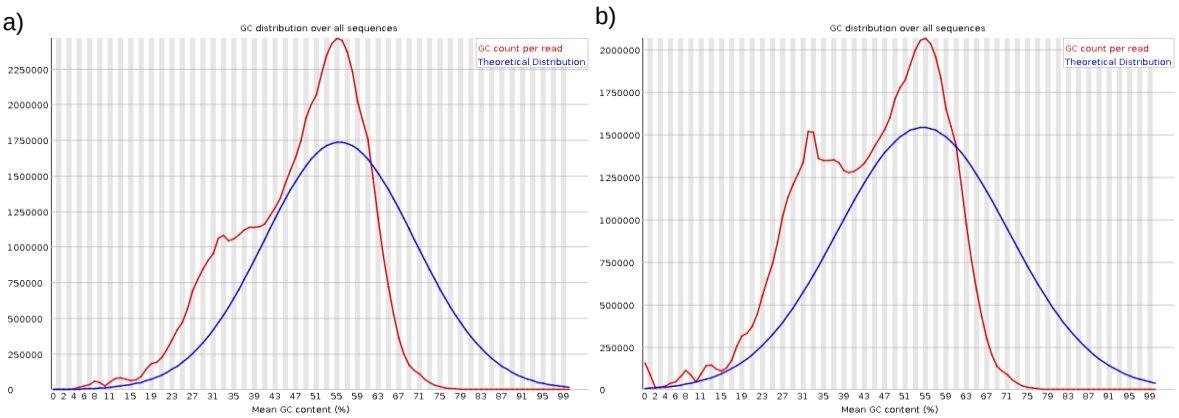


Figure 18 – Read GC content distribution for a) WT and b) X7/C24. The blue line represents the theoretical distribution curve, whereas the red line represents the actual GC count content found on the libraries.

Focusing on the abnormal peak formed in the X7/C24 libraries, the mean read distribution between 19 and 40% total GC content was as follows: 95% of the reads on the interval were aligned to the genome in both WT and X7C24, and 68,8% of those had a gene

feature in the X7/C24 dataset, versus 69,8% in the WT. These results dismiss the possibility of contamination from other species, as the remainder 5% of non-aligned reads would not cover for the observed read increase.

To assess for retained introns, we searched within the reads with 19-40% GC contents for reads that overlap with known *Drosophila* introns. Results showed that the ratio between reads that featured introns and reads that featured protein coding genes was constant between conditions and GC content percentage (Figure 19) leading us to the conclusion that the extra peak in the 19-40% GC content area was not caused by an increase of intron retention, as no significant changes were being observed across conditions.

Some studies suggest the differences in GC content distribution are related to library preparation, especially due to PCR duplication[58]. Considering PCR duplication accounts for an average of 30% of the datasets studied here, we believe that the origin of the peak may be related to library preparation. As mentioned before, results also showed that this peak did not seem to be affecting the alignment results, suggesting the downstream analysis would not be impaired.

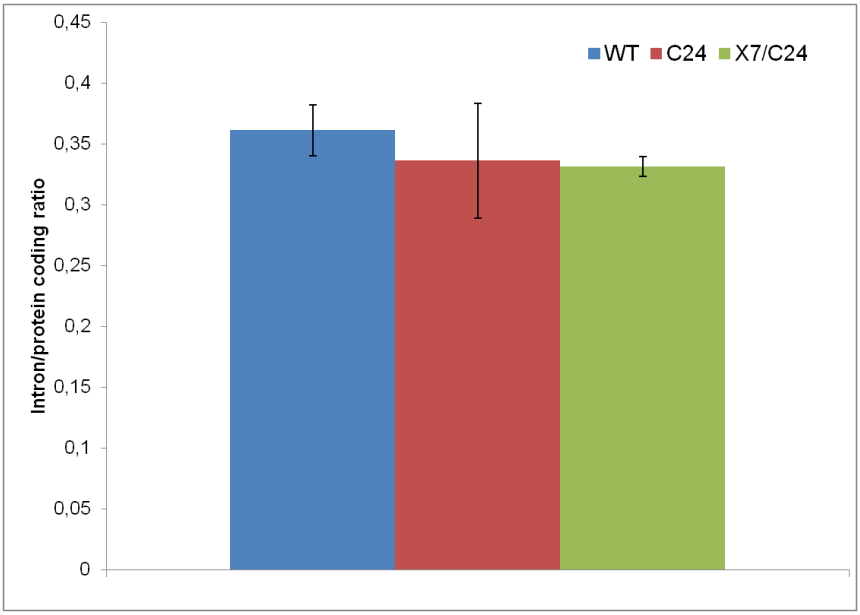


Figure 19 – Intron feature/protein coding feature ratio for the second and third sequencing batch libraries on reads with 19-40% of GC content.

In order to identify differences in gene expression values in the X7/C24 libraries, we compared them to the WT libraries of the second sequencing batch via a DEA with DESeq. Results show a list of 2844 DE genes (adj-p < 0,05), including the hallmark gene *Smn* ($\log_2$ fold change: -1,15) as expected. To assess which pathways were being affected by the gene expression changes found, we performed a GSEA for the X7/C24 DE gene list. Results show 393 significantly enriched biological process (BP) pathways (adj-p < 0,05) with an emphasis on pathways related to neuronal processes, splicing, RNA processing and protein assembly (Table 3), all of which are results that are in agreement with the functions attributed to *Smn*

and its down-regulation, as well as SMA.

Table 3 – Selected terms (lowest adj-p) from the GSEA (BP) for the list of genes obtained in the WTxX7/C24 gene DEA. All terms with a term size higher than 500 were filtered out due to being too high up in the ontology.

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0048666	232	420	neuron development	7,54E-23
GO:0000904	235	427	cell morphogenesis involved in differentiation	5,66E-22
GO:0007268	139	225	synaptic transmission	3,89E-19
GO:0051960	96	146	regulation of nervous system development	6,84E-16
GO:0007411	106	187	axon guidance	1,90E-11
GO:0048523	191	422	negative regulation of cellular process	2,36E-09
GO:0016071	147	300	mRNA metabolic process	3,42E-09
GO:1901861	44	65	regulation of muscle tissue development	2,15E-08
GO:0007417	110	215	central nervous system development	2,17E-08
GO:0006396	171	379	RNA processing	2,16E-07
GO:0007269	37	55	neurotransmitter secretion	2,82E-07
GO:0045664	37	56	regulation of neuron differentiation	5,95E-07
GO:0000375	110	227	RNA splicing, via transesterification reactions	6,20E-07
GO:0050768	23	29	negative regulation of neurogenesis	7,76E-07
GO:0031644	34	52	regulation of neurological system process	2,51E-06
GO:0001558	31	46	regulation of cell growth	2,78E-06
GO:0050804	33	51	regulation of synaptic transmission	4,73E-06
GO:0016079	29	43	synaptic vesicle exocytosis	5,34E-06
GO:0034329	30	45	cell junction assembly	5,39E-06
GO:0000398	99	212	mRNA splicing, via spliceosome	1,27E-05

By separating the DE gene list between up-regulated genes and down regulated genes in and performing a GSEA in order to see if there were specific pathways with mainly negative or positive changes, we found the same previous pathways significantly enriched distributed between both analyses. We also observed a clear separation of types of pathways between up and down-regulated genes. The up-regulated genes showed an enrichment of pathways related with several growth differentiation-related processes, including the morphogenesis of brain, neurons and eye, neuron projection and muscle development as well as mRNA splice site selection-related genes (see Annex III – *D. melanogaster*). On the other hand, the negatively regulated genes show a down-regulation of ribosome biogenesis, protein synthesis, neuronal development, splicing and DNA repair. Upon further observation, 1 out of the 31 genes found as having a ribosome biogenesis function was also described as a regulator of alternative splicing[59], while 17 other genes were described in a study as also related to neurogenesis[60]. SMN down-regulation, while commonly associated to neurogenesis-related processes, such as brain development[61] and neuronal migration/differentiation[62], has recently been found to have a function within the translation process[63]. Results were consistent with this study, as we observed an enrichment

of pathways related to protein synthesis and ribosome biogenesis.

Next, we investigated if the decrease of SMN expression had an impact on the splicing events occurring in the CNS. For this purpose we used DEXSeq, a software that tests for differential exon usage between two conditions. The analysis found 8304 exons as differentially expressed, corresponding to changes in the isoform expression of 3926 genes, 1127 (28,7%) of which were also listed as DE genes. We also performed a GSEA for the list of X7/C24 genes with differential exon usage. This analysis showed (see Annex III – *D. melanogaster*) an emphasis on terms related to RNA splicing, neurogenesis, axon guidance, post transcriptional gene silencing and assembly of the spliceosomal complex. Focusing on the 33 genes associated to regulation of alternative mRNA splicing, via spliceosome, we found these genes were mostly related to the assembly of snRNP's (mainly U2 and U12). Considering that SMN down-regulation has been previously shown to perturb the splicing of U12 intron-containing genes in vertebrates and *Drosophila*[64], we investigated if our results were consistent with these findings. However, we could not confirm if the specific U12 introns described in this study were being affected, only the genes which contain these introns. Out of the 24 genes listed with U12 introns, 9 of them were found as having DE exons in our analysis, totalling 23 DE exons (Figure 20). 4 of the genes with predicted U12 intron targets were also found on the X7/C24 gene DEA - Tsp97E, Nhe3, CG16941 and CG17912. From these, only CG16941 has been shown to affect the regulation of alternative mRNA splicing[59]. These results, while not showing if the specific U12 introns are being affected, show that changes in SMN expression affect genes modified by U12.

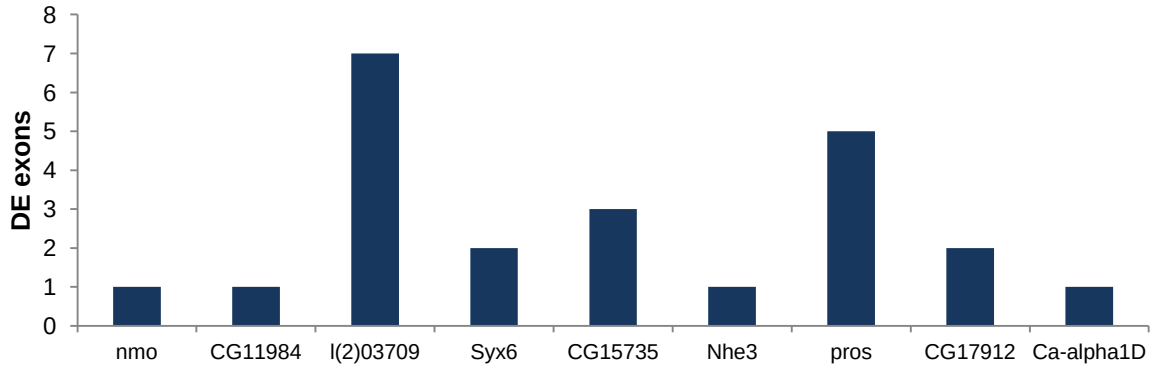


Figure 20 – Number of DE exons found on each gene predicted to be modified by the U12 snRNP.

Finally, in addition to the pipeline described in the methods, we also used the Tuxedo pipeline[65] to map and quantify gene and isoform expression, as it is one of the most widespread analysis pipelines in bioinformatics and it provides the feature of detecting novel splicing isoforms, something that is not implemented in DEXSeq. This pipeline uses TopHat[15] for mapping reads, Cufflinks[66] for transcript assembly, and Cuffdiff [65] for gene and isoform DEA. Results show that Cufflinks is not optimized for novel isoform discovery in *Drosophila*, resulting in a systematic prediction of gene fusion transcripts (Figure 21). This is probably

due to the difference in genome architecture between mammals and fly, which greatly differ in intron size[67]. While it is possible to change the average intron size and intergenic distance in TopHat's options, some intergenic distances in *D. melanogaster* are as long as its introns and are not differentiated by Cufflinks. For example, *Smn*'s distance to *nxf2* is less than 1kb, leading to a systematic fusion of both genes into a single transcript.

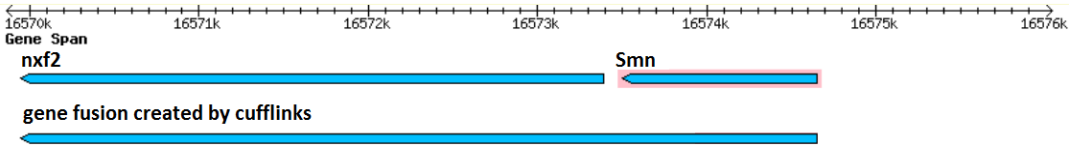


Figure 21 – Example of aberrant transcripts predicted by the Cufflinks algorithm resulting from an artificial fusion of sequencing reads from closely positioned genes.

3.1.4 – Assessment of the effect of read trimming on nucleotide frequency bias and read coverage

We observed that the first 10 reads on all libraries across all conditions show a bias in the per base sequence content (Figure 22), which has been previously described as a result from the use of random hexamer primers[4] in library preparation. A second data analysis was performed, with the first 10 nucleotides trimmed off in order to determine if it would reduce the bias, provide higher alignment quality and consequentially a more accurate gene and exon DEA when compared to the full length read analysis.

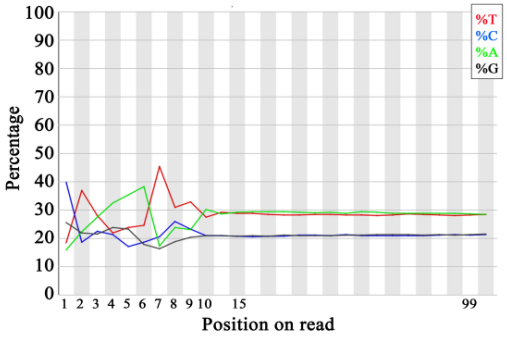


Figure 22 – Nucleotide abundance across read positions.

Results showed that trimming reads reduced the library size after the initial filtering steps (homopolymers, poly-A, and read quality), but increased the number of uniquely mapped reads to the genome (Figure 23a). A decrease in introns was also observed, due to the fact that since HTseq requires only a partial overlap between the coordinates of read positioning and gene annotation in the reference genome, it is more than likely that the number of partial overlaps will be reduced. Results also showed that the increase in aligned reads does not help in stabilizing variance between replicates. A gain in reads corresponding

to protein coding feature was also observed, increasing coverage in medium and long genes (Figure 23b). This gain in protein coding feature genes also caused changes in the DE analysis, with a small number of unique genes being found on both trimmed and full read approach and an overall decrease in DE genes (Figure 23c and 23d). Even though trimming the first 10 nucleotides somewhat improves read mapping, we found diminishing results in the downstream analysis when performing the DEA. Because of this and time constraints, we decided not to pursue the complete *Drosophila* analysis with read trimming, though if the same base sequence bias were to be observed in another study, we would recommend using the trimmed approach from the start.

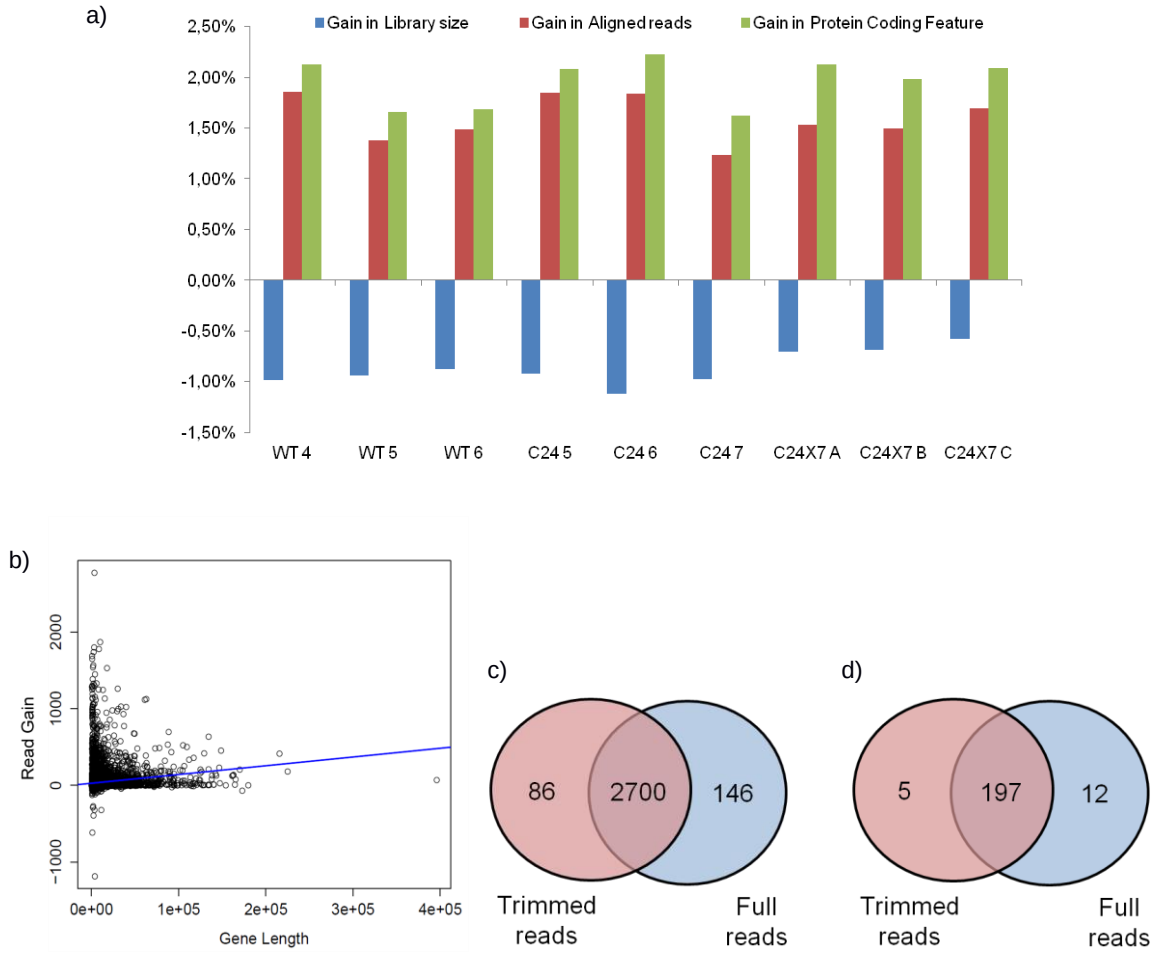


Figure 23 – Aligned read gain by trimming the first 10 nucleotides of each read. a) Gain in aligned reads, library size and reads with a protein coding feature for each condition. b) Linear regression of the number of reads gained across gene length. c) Overlap between Trimmed and Full gene DEA for X7/C24. d) Gene overlap between Trimmed and Full gene DEA for C24

3.1.5 – Comparative analysis of the transcriptome profiles of C24 and X7/C24 flies

Briefly, the creation of the C24 model had the objective of identifying neuronal specific changes caused by the down-regulation of SMN in the CNS. However, results showed this approach was hindered due to library preparation problems and also lead us to hypothesize

that the presence of glial cells in the CNS, which have a high SMN expression, diluted the effects of the shRNA construct since these cells do not express elav. To confirm *Smn* expression levels in these libraries, a RT-qPCR quantification from total RNA samples was performed which showed a non-significant down-regulation of SMN (Figure 24), in agreement with the quantification found in the RNA-Seq libraries. We then expressed the same shRNA construct in a model which presents a genomic deletion of *Smn*, halving its expression in all tissues. This approach, while not as precise in finding neuronal-specific changes, produced a wider list of results which are consistent with previous publications in terms of affected pathways. Also, the RT-qPCR quantification of *Smn* for these libraries showed a very significant down-regulation (Figure 24).

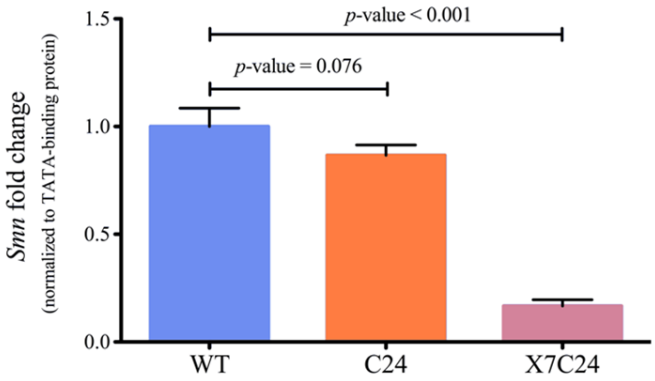


Figure 24 – RT-qPCR of SMN expression levels. From Amaral et al.(2013)[41]

Comparing gene DEA results between these two models, we found that 61 of the 79 (77%) DE genes found in the 3x3 C24 DEA overlapped with the X7/C24 DE gene list, and 152 out of the 209 (73%) DE genes from the 2x2 DEA overlapped with the X7/C24 DE gene list. In total, 49 genes were found in common between all three DEAs (Figure 25a). A number of the X7/C24 DE genes were also found in two other publications on SMA based on *Drosophila* models using similar fly disease models, one using *Smn* null larvae sequenced with RNA-Seq (Garcia et al.[68]), and the other based on a microarray study of a loss of function allele for *Smn* (*Smn*^{73Ao} - Lee et al[69]). We found an overlap of 12 genes (Figure 25b), including the hallmark gene *Smn*. The gene *cer* (or *crammer*) was also found in all three studies and is known to be involved in long term memory regulation[70]. Although no relation between it and neuro-muscular junctions has been reported, a link between *cer* and another neurodegenerative disease, Alzheimer disease, has been made[71]. The other genes – *CG10638*, *CG2233*, *CG6910*, *CG7394*, *CG9577*, *Cht5*, *d*, *e*, *RhoGEF4*, and *Spn43Aa* – did not show a direct relation with splicing, motor degeneration, or neurodegenerative diseases.

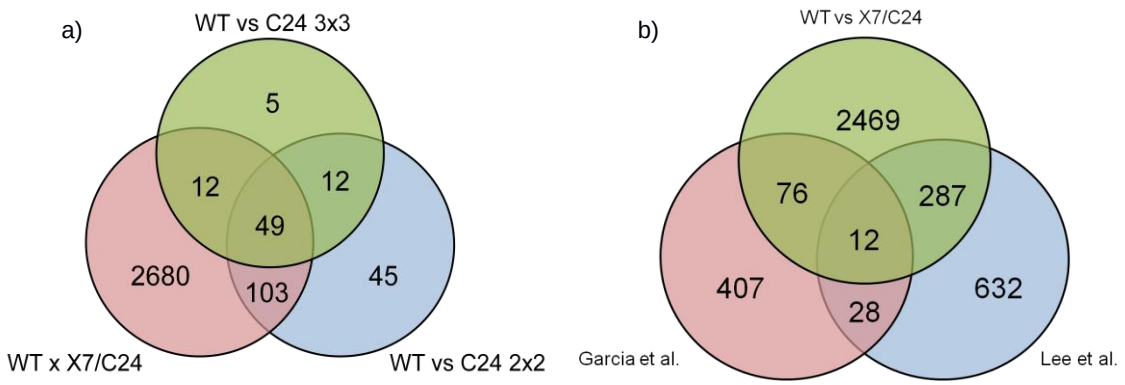


Figure 25 – Overlap between the differentially expressed genes found in C24 and X7/C24 libraries. a) Overlap of the genes found in the two DEAs performed for C24 and the DEA for X7/C24. b) Overlap of the genes found in two previous studies and the DEA for X7/C24 libraries. The study represented by the red circle is based on RNA-Seq data (Garcia et al.[68]) while the other, represented by a blue circle, is based on microarrays data (Lee et al.[69])

As previously discussed, we established the hypothesis that the glial cells could be interfering with the effects of the shRNA in the down-regulation of SMN's expression levels, which was then reinforced by the observation that more than 70% of the significant DE genes found on each of the C24 DEAs were also found on the X7/C24 DEA (Figure 25a). This suggested a change in the gene expression pattern that is dependent of SMN's expression, despite the gene itself not being found as significantly DE, though it was not possible to test this with the C24 data alone. However, with the data produced in the X7 libraries we were able to perform a two-proportion z-test and determine if the impact of glial cells was statistically significant. To test this, we retrieved all genes annotated in Flybase as expressed in "neuron", "glia" and "ubiquitous" and overlapped with the the gene DEA results. As seen below (Table 3), we found that there is a significantly smaller proportion of ubiquitous genes significantly DE in the WT vs C24 list, while having a higher sensitivity for expression changes in neuronal specific genes. This helps to confirm the established hypothesis, showing that even though there is no significant change in the expression of SMN in the WT vs C24 analysis, we can attribute this to a signal dilution effect caused by cell types not affected by the shRNA, as the effects of a lower SMN expression level are still being observed.

Table 4 - Differentially expressed genes in C24 or X7/C24 flies that are classified as having neuronal, glial or ubiquitous expression in Flybase. The (*) symbol represents statistically significant differences in the relative proportion of identified genes between the two sample types using the two-proportion z-test are highlighted with ($p < 0,05$).

	Neuronal (442 genes)	**Glial (74 genes)**	**Ubiquitous (2649 genes)**
C24 (209 DE genes)	20 genes* (9,5%)	1 gene (0,5%)	39 genes* (18,5%)
X7/C24 (2846 DE genes)	109 genes (3,8%)	29 genes (1,0%)	743 genes (26.1%)

3.2 – Transcriptome profiling of motor neurons derived from SMA patient iPSCs

To study the effects of SMN's down-regulation on human motor neurons differentiated from iPSC cultures, we analysed three different phenotypes. This analysis had the aim of uncovering which genes, exons and related pathways are affected by this down-regulation, as well assessing if the previous studies, made with non-human models, showed a high degree of conservation regarding ortholog gene expression changes. Like the *Drosophila* analysis, it followed the pipeline described in methods, which includes data quality filtering of the generated datasets, alignment to a reference genome, differential expression analysis of genes and exons, and gene set enrichment analyses. Also, given the results obtained from trimming the first 10 nucleotides on *D. melanogaster* were successful in reducing bias and raising protein coding gene coverage we decided to pursue this approach in the human data analysis from the start.

The human RNA-seq dataset consisted of 9 libraries with three replicates for three different conditions: control motor neuron cultures derived from a normal subject (NS), motor neuron cultures from the same cells transduced with an RNAi lentiviral vector targeting SMN1 (shSMN2) and motor neuron cultures derived from a type I SMA patient (SMAiPS). While the NS and shSMN2 datasets are composed of biological replicates, two of the SMAiPS libraries are technical replicates (SMAiPS 2 and 3) derived from the same pool of extracted RNA. Since there were only two individuals used for the generation of the iPSC cultures (one for the normal subject and one for the patient), the shSMN2 condition was developed to address the inherent genetic background variation between individuals.

An average of 101 million raw reads per library were obtained (Appendix I – *H. sapiens*), 90% of which passed the initial quality filtering process. On average, 84,5% of the filtered reads were aligned to the human genome and 58% of them were counted as part of a protein coding gene. After filtering, 79% of the transcriptome was covered, with an average sequencing depth of 79X and an average of 19469 protein coding genes expressed, out of a total of 25775. The filtering, trimming and duplicate removal did not discard any of the replicates on either approach (Figure 26a) however, since SMAiPS 2 and 3 are technical replicates the library with the lower correlation value to SMAiPS1 was removed from the downstream analysis ($\rho_{SMAiPS\ 2}$ = 0,987 vs $\rho_{SMAiPS\ 3}$ = 0,988 in the correlation test). The three conditions showed a distinct phenotype assignment, separating the libraries according to their origin (Figure 28b).

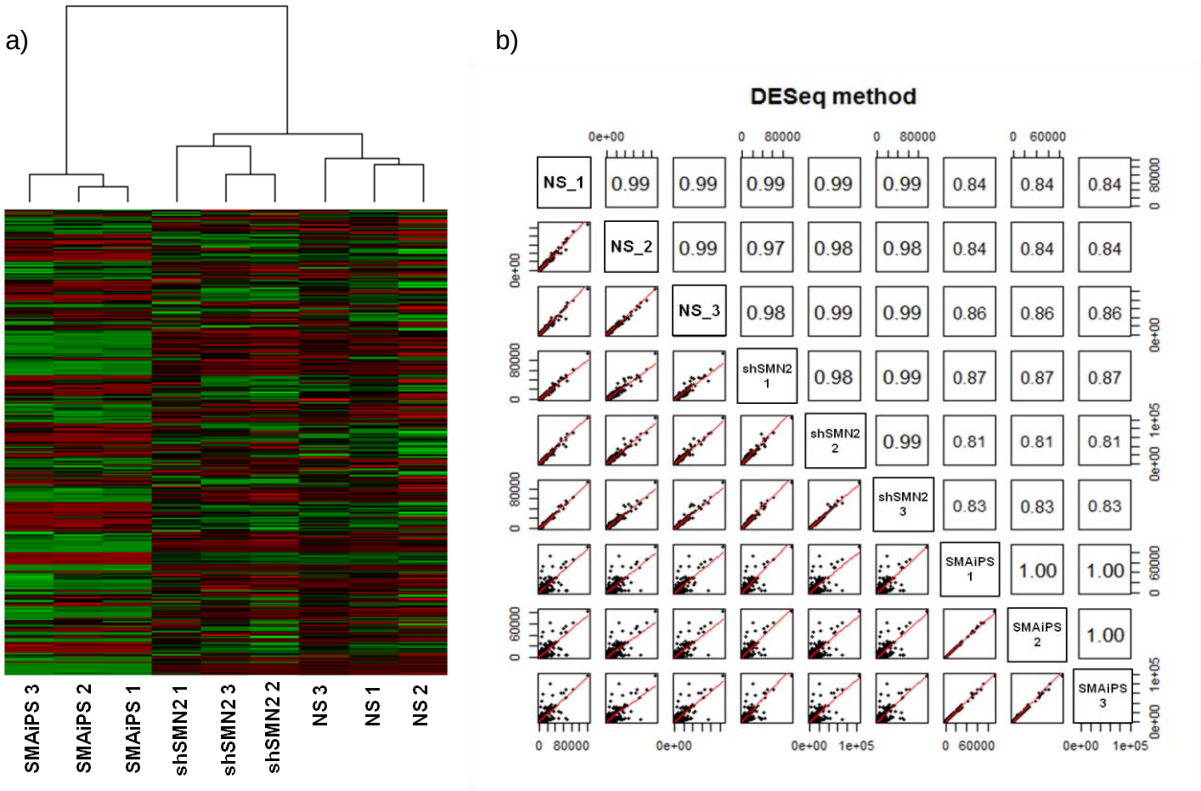


Figure 26 – Human data assessment. a) Hierarchical clustering between human, showing a separation between all three conditions. b) Correlation plot between human libraries based on gene expression values normalized by DESeq.

To assess which genes were having its expression levels altered by SMNs down-regulation in the SMA patient and the shSMN2 model, we performed a gene DEA. Two comparisons were made: NS vs SMAiPS and NS vs shSMN2. In the NS vs SMAiPS DEA, we found 8209 genes as DE (adj-p < 0,05) including SMN1 (log_2 fold change: -7,55) and SMN2 (log_2 fold change: 1,12). Many of these genes displayed medium to low expression levels which suggests the DE analysis might be biased due to sampling error. Comparatively, we found a much lower number of DE genes in the NS vs shSMN2 DEA (208 DE genes) and it did not feature either *SMN1* or *SMN2* as DE.

As we did not find significant expression changes of *SMN1* in the shSMN2 libraries, we decided to perform a DEA between the two libraries which were expected to have a down-regulation of SMN – shSMN2 and SMAiPS. Results showed 8101 DE genes, 6803 of which overlapped with the NS vs SMAiPS analysis (~84% - Figure 27), including a down-regulation of SMN1 (log_2 fold change: -7,21) and an up-regulation of SMN2 (log_2 fold change: 1,09) with practically the same fold changes as the one found in the NS vs SMAiPS DEA. Since the shSMN2 construct was developed to solve the problem regarding the genetic background between iPSC cultures, this points to a problem with this model, probably due to the shRNA interference being weak. This could be related to an immune response by the cells to the shRNA (previously documented in mammalian cells[72]), disabling any effects

caused by RNA interference.

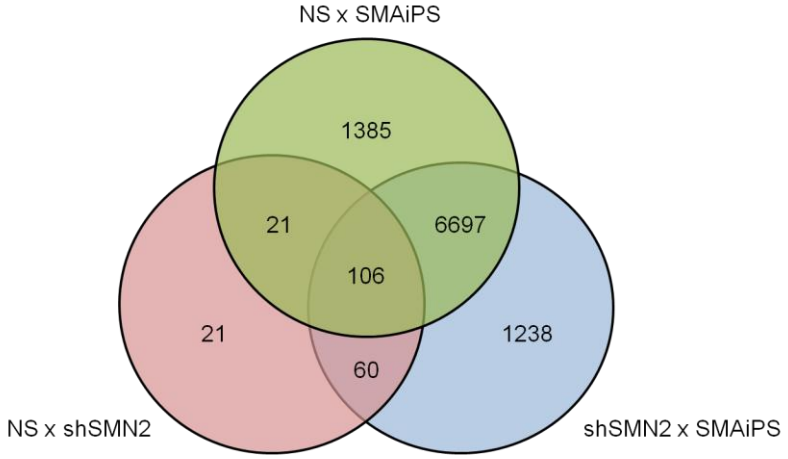


Figure 27 - Overlap between the genes found as DE in the *H. sapiens* sequencing batch.

To further understand the biological context of these gene expression changes, we assessed which pathways were being affected by these variations by performing a GSEA. We found 356 significantly enriched BP pathways with the DE gene list from the NS vs SMAiPS analysis, 6 from the NS vs shSMN2 and 323 from the shSMN2 vs SMAiPS (adj-p < 0,05). In both NS vs SMAiPS and NS vs shSMN2 GSEAs, we observed a high number of genes associated with neuronal growth and activity, and axon guidance (Table 5 and 6), which are biological processes that have been described in the literature as negatively affected by the down regulation of SMN.

Table 5 - Selected terms (lowest adj-p) from the GSEA (BP) for the gene list obtained in the NSxSMAiPS gene DEA. Terms with a term size higher than 500 were filtered out. Other relevant terms not shown here include central nervous system projection neuron axonogenesis, neuron maturation, neuron projection morphogenesis and central nervous system neuron development

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0051960	289	441	regulation of nervous system development	1,70E-15
GO:0045664	176	258	regulation of neuron differentiation	6,53E-12
GO:0007411	228	354	axon guidance	3,26E-11
GO:0007155	223	352	cell adhesion	6,60E-11
GO:0007268	246	392	synaptic transmission	8,19E-11
GO:0007610	264	429	behavior	4,57E-10
GO:0030198	137	199	extracellular matrix organization	7,92E-10
GO:0001568	287	477	blood vessel development	2,78E-09
GO:0034329	77	101	cell junction assembly	5,22E-09
GO:0001525	214	346	angiogenesis	1,71E-08
GO:0051240	271	454	positive regulation of multicellular organismal process	1,71E-08
GO:0060284	208	337	regulation of cell development	2,22E-08
GO:0003001	199	320	generation of a signal involved in cell-cell signaling	3,03E-08
GO:0022604	187	299	regulation of cell morphogenesis	4,42E-08
GO:0042060	288	491	wound healing	4,90E-08

Table 6 - Selected terms (lowest adj-p) from the GSEA (BP) for the gene list obtained in the shSMN2xSMAiPS gene DEA. All terms with size higher than 500 were filtered.

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0051960	269	406	regulation of nervous system development	1,52E-15
GO:0031344	180	256	regulation of cell projection organization	3,95E-14
GO:0045664	215	318	regulation of neuron differentiation	6,07E-14
GO:0022604	194	297	regulation of cell morphogenesis	1,39E-10
GO:0007411	224	352	axon guidance	1,63E-10
GO:0007155	265	439	cell adhesion	1,09E-09
GO:0051240	274	451	positive regulation of multicellular organismal process	1,10E-09
GO:0001568	282	467	blood vessel development	1,64E-09
GO:0007517	201	321	muscle organ development	1,03E-08
GO:0032970	132	199	regulation of actin filament-based process	4,50E-08
GO:0001501	156	243	skeletal system development	4,83E-08
GO:0090066	172	273	regulation of anatomical structure size	7,85E-08
GO:0007268	226	378	synaptic transmission	1,25E-07
GO:0010740	197	322	positive regulation of intracellular protein kinase cascade	1,30E-07
GO:0050920	70	94	regulation of chemotaxis	1,39E-07

Table 7 – Enriched terms from the GSEA (BP) for the gene list obtained in the NSxshSMN2 gene DEA.

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0060337	6	59	type I interferon-mediated signaling pathway	0,000319376
GO:0034340	6	60	response to type I interferon	0,000319376
GO:0060282	2	2	positive regulation of oocyte development	0,000319376
GO:0032836	3	10	glomerular basement membrane development	0,000337125
GO:0016525	5	47	negative regulation of angiogenesis	0,000368576
GO:0060384	3	14	innervation	0,000656759

The results for the NS vs shSMN2 evidenced an enrichment of pathways associated to interferon-based immune responses (Table 7). The two first terms were also found on the NS vs SMAiPS GSEA and in both of these analysis, all genes related to them were found to be down-regulated.

We also found a high number of genes related to viral infection such as the Hepatitis C disease pathway (KEGG term) when comparing the healthy individual's iPSC cultures against the SMA patient (79 genes for NS vs SMAiPS, 76 for shSMN2 vs SMAiPS, out of a total of 116 known genes in this pathway. See Appendix III – *H. sapiens*). It is possible that the healthy individual fibroblasts contained a virus infection which carried over to the iPS cell cultures, which would explain the down-regulation of genes related to immune response in the SMAiPS libraries when compared to NS.

Finally, we assessed if the decrease of *SMN1* expression causes significant changes in the cell genetic program in terms of differential exon usage. Results showed (adj-p <0,05) 3052 DE exons in the NS vs SMAiPS analysis, corresponding to changes in 1496 genes,

4864 DE exons in shSMN2 vs SMAiPS (2254 genes), and finally 1991 exons in the NS vs shSMN2 analysis (1324 genes) (Figure 28).

Unlike in the GSEA for the DE gene list, the GSEA for NS vs SMAiPS DE exons showed 53 genes (see appendix III – *H. sapiens*) directly associated with RNA splicing including SNRNP200 (required in U4/U6 snRNP assembly), SNRPB2 (related to U2 snRNP) and SNRNP70 (encodes U1 snRNP). Pathways enriched in NS vs shSMN2 are again mostly related to virus-host interactions like the ones observed on DE gene GSEA.

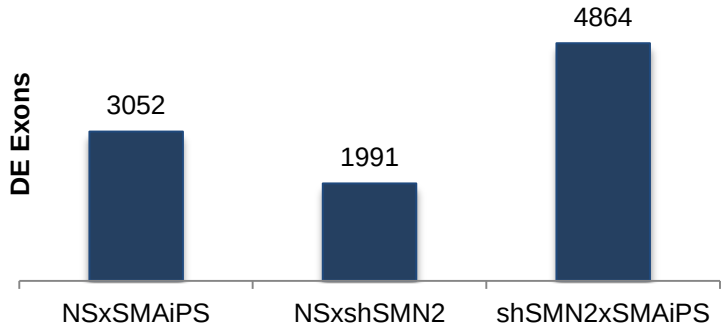


Figure 28 – Differentially expressed exons found on the human library DEXSeq analysis.

These observations, coupled with the fact that 90% of the genes in the shSMN2 vs SMAiPS gene DEA overlap with NS vs SMAiPS's DEA, and the much closer correlation between the NS and shSMN2 libraries (~99%, see Figure 26b) versus the SMAiPS libraries (~85% and ~84%, respectively), lead us to the conclusion that the shSMN2 model was indeed not performing as expected, as there was no observable down-regulation of SMN and the gene DEA versus SMAiPS showed that there were no significant differences between it and the DEA between NS and SMAiPS.

Given the results presented here, we decided to compare the data to previous SMA studies to see if the same type of expression changes and pathways were found between studies. The caveat with this approach is that since most studies on SMA have been made on non-human models only known ortholog genes will be picked up by a comparative analysis, potentially discarding a long list of genes that have not previously been associated with SMA or are specific to a certain species. We used two mouse SMA studies based on microarrays for this comparison. The first study, by Zhang et al.[73], is based on both brain and spinal cord tissues and uses shRNA to limit the expression levels of SMN in the mouse, similar to the ones presented on this thesis, with the exception that the interference in this case is ubiquitous. The second study, by Bäumer et al.[36], is based on a spinal cord tissue assessment of a SMN null mouse model which expresses human *SMN2* and SMNΔ7, known to extend the life-span of mouse SMN models (previously described in[74]). Results show a high overlap between DE human genes and DE mouse genes (~50% each), though a very small overlap was observed when overlapping the mice studies (~7% overlap) (Figure 29).

To confirm if these overlaps were statistically significant, we did a hypergeometric test using the total number of ortholog genes between mouse and human as the universe. Results showed the overlap made with Zhang et al. is statistically significant (p-value=0.032), while the overlap with Bäumer et al. is not (p-value=0.351).

We performed two GSEA analyses for the set of overlapping genes between our and these models. The GSEA for the Zhang et al. overlap found 50 significantly enriched biological process terms (see appendix III – Human-Mouse orthologs). The results are consistent with reported changes associated to the down-regulation of SMN expression, including neurogenesis, axon guidance, and neuron differentiation. No terms related to splicing were detected in the GSEA. Additionally, pathways such as viral infection, which had been attributed in humans to the biological variation between individuals, were not found.

On the other hand, the GSEA for the Baumer et al. overlap showed 19 enriched terms (see appendix III – Human-Mouse orthologs), which are not known to be reported as associated with SMA or SMN down-regulation. Our results suggest this model may not be appropriate to model SMA, as the enriched biological process terms found in the GSEA do not mirror the known SMA phenotype. We hypothesize this could be related to the fact that while SMNΔ7 is unstable in humans and is rapidly degraded, its effects on mice cause changes in gene expression unrelated to the SMA pathology.

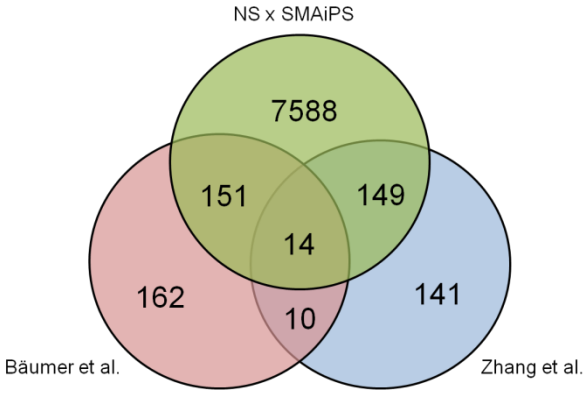


Figure 29 – Overlap between DE genes in the NSxSMAiPS DEA and two SMA studies on a *M. musculus* model.

3.3 – Integrated analysis of human and fly SMA models

To understand how conserved were the gene expression changes found in *Drosophila* in comparison to humans, we overlapped the X7/C24 DE gene list with the list of all known human-*Drosophila* orthologs. The same was made for the NS vs SMAiPS DE gene list. The two overlaps were then compared. We found 1419 gene hits in common between the human and *Drosophila* gene DEAs. A gene hit means that one gene may be an ortholog to more than one human/*Drosophila* gene (Figure 30a). To assess if the overlap was statistically significant, we performed a hypergeometric test. Results show that this overlap is significant (p-value: $2.59e^{-14}$). Plotting the fold changes of these genes between species, we obtained a list of 740 genes (green) with the same type of expression change (i.e.: both up-regulated or down-regulated - Figure 30b).

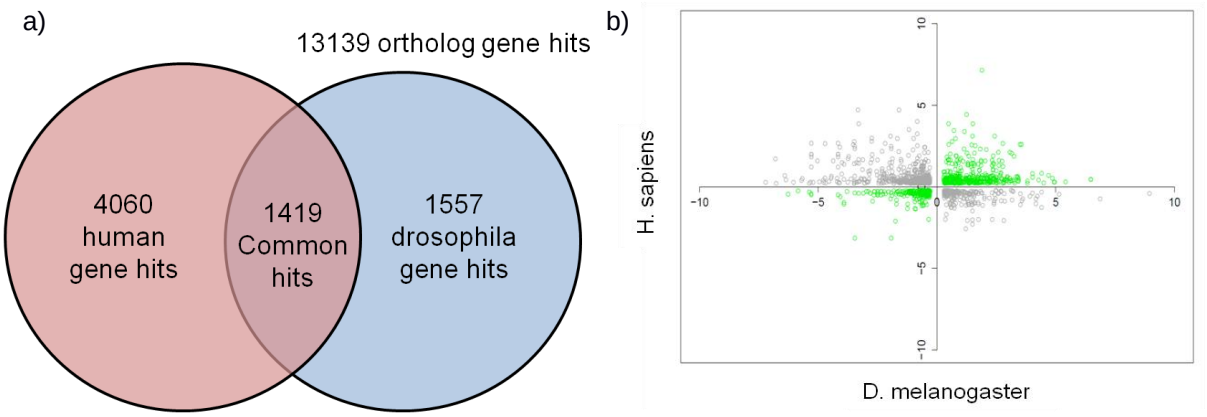


Figure 30 – *H. sapiens* and *D. melanogaster* differentially expressed genes which are also orthologs. a) Venn diagram representing the gene hits found in common between the DE gene lists for *H. sapiens* and *D. melanogaster* which are also ortholog genes between these species. b) Comparison between the fold changes of the ortholog genes found to be differentially expressed in both *H. sapiens* and *D. melanogaster*. The circles in green represent genes whose fold change is the same in both species (both positive /both negative).

For the list of 740 genes in common we performed two GSEAs, one for the *Drosophila* gene list and one for the human gene list, in order to assess if the ortholog genes were involved in similar pathways. The *Drosophila* GSEA found a set of 156 enriched biological process terms (adj-p < 0,05), with an emphasis on neuron related processes and transportation, 68 of which are consistent with the ones found on the GSEA for the X7/C24 DE genes. The human GSEA found a set of 51 enriched biological process terms (adj-p < 0,05) with an emphasis on neuron related processes, RNA processing, ribosome biogenesis and protein assembly (see appendix III – *Drosophila*-human orthologs).

As mentioned in the *Drosophila* model GSEA analysis, the biological pathways related to ribosome biogenesis and protein assembly are in agreement with a recent find of a new SMN function related to translation[63]. Sanchez et al. found that in *in vitro* cultured human cells SMN associates with polyribosomes and is able to repress the translation of mRNAs. Mutations in SMN thus result in the upregulation of these mRNAs in SMA patients.

By showing an enrichment of genes involved in these processes, our results corroborate the suggestion that these specific pathways are affected by mutations in the SMN protein. Furthermore, these results suggest the comparison method is able identify genes and pathway changes which are central in the SMA pathology, filtering genetic variation related noise in the data. Despite the limitations discussed previously for both datasets, the results presented here give evidence for a conserved response to the down-regulation of SMN between species, with a significant overlap between the affected genes and pathways, suggesting the *Drosophila* model is indeed a good approach to model and study SMA disease.

4 – Final Remarks

The objectives in this thesis were to investigate how the decrease of SMN expression affects the genetic program of motor neurons via an RNA-seq approach. To do this, we studied the transcriptome of two disease models using RNA-Seq: a *D. melanogaster* RNAi model, and a human iPSC model of an SMA patient, both of them with decreased levels of SMN expression. We also aimed to assess the similarity between *Drosophila* and human models in order to understand the usability of *D. melanogaster* models in studying SMA.

Regarding the *Drosophila* model, results presented here are the first map of the transcriptome of the CNS of a *Drosophila elavGAl4* transgenic strain, which contains a neuronal specific expression of a shRNA construct that targets the expression of *Smn*. Overall, across all three sequencing batches, we found that the achieved alignment coverage was consistent with other sequencing datasets on *D. melanogaster*, comparable to a mRNA-seq *Drosophila* brain study by Hughes et al.[75] and two whole-body transcriptome studies by Daines et al. and Gan et al., respectively[76,77]. Similarly to Hughes et al., only 70% of the reads generated in this dataset map to protein coding genes, 20% less than the modENCODE study for whole-body sequencing, suggesting that the *D. melanogaster*'s CNS transcriptome could potentially be enriched with intergenic transcripts, as reported in primates[78]. One of the biggest challenges found on this model was the amount of non-neuronal tissue found on all datasets, due to the larvae brain extraction process needed to produce the libraries. Since RNA-Seq requires a large amount of RNA input, each library was comprised of an average of 200 larvae brains, each of them with varying levels of contamination from non-neuronal tissues. As mentioned above, we successfully developed a method to determine tissue specific expression and the effects of tissue contamination on library viability. This method and other solutions discussed here resulted in a publication while this thesis was being written[41], focusing exclusively on the *D. melanogaster* shRNA model. Also, as shown in the results, RNA-Seq, while a viable and flexible technique, has various limitations related to library preparation and sequencing. A great part of the challenges found on these libraries was derived from library preparation, especially PCR amplification, which consequently forced discarding an average of 50% of the dataset. Finally, our results suggest that RNA-seq, for a proper comparison between different conditions, requires libraries that are prepared and sequenced in parallel, therefore mitigating the library preparation and sequencing batch effects.

Likewise, we found several challenges with the human model analysis. The approach used to take into account the biological variation existent between two individuals was to create a model based on the healthy individual where the SMN expression was decreased due to the targeting of an SMN specific shRNA. Results showed however, that this approach

did not work, leaving us with two libraries displaying a high gene expression variance, mostly due to the inherent sampling error of the technique that mainly affects medium to low expressed genes. Indeed, we found almost a third of the human annotated genes as differentially expressed when comparing them (NS vs SMAiPS). Still, we managed to determine which genes were potentially related to a response to SMN's down-regulation by comparing those results to previous mouse-based SMA studies.

Despite having problems in both *Drosophila* and human models, it was possible to compare the gene expression between them, revealing several genes and pathways related to the decrease of expression conserved between species, indicating that the *D. melanogaster* is an apt model for studying SMA.

The analysis presented here sets up a starting point for a future, more accurate transcriptome study in both models using RNA-Seq. Regarding the *D. melanogaster* model, improving the sequencing libraries quality is key, which involves reducing the number of PCR cycles during library preparation (thus resulting in fewer duplicates). Also, for an unbiased analysis of the *Drosophila* data, a new dataset containing new libraries for WT and X7/C24 will be required, which will also allow for a DEA that can take into account a possible sequencing batch effect when compared to the previous sequencing batches. Future shRNA based *D. melanogaster* models will also need to address the fact that since the CNS is a complex tissue, the presence of glial cells can undermine the process of detecting subtle variations in gene expression in the CNS, especially if the shRNA expression is mediated by *elav,* which is not expressed in these cells. Ideally, the best model would be based on completely isolating the neurons from the rest of the CNS. However this is an extremely technically complex procedure for it to be used in *Drosophila*. As for the *H. sapiens* iPSC model, sequencing batches need to be improved by increasing the number of biological replicates combined with the increase of sequencing depth, specifically by creating iPSC cultures from several different individuals, rather than using only one for each condition, therefore mitigating the false discovery rate by reducing the genetic background and increasing sequencing sampling.

5 – Bibliography

1. Twine, N. a, Janitz, K., Wilkins, M. R. & Janitz, M. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PLoS One* **6,** e16266 (2011).

2. Cooper-Knock, J. *et al.* Gene expression profiling in human neurodegenerative disease. *Nat. Rev. Neurol.* **8,** 518–30 (2012).

3. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

4. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38,** e131 (2010).

5. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12,** R22 (2011).

6. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21,** 1543–51 (2011).

7. Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9,** 387–402 (2008).

8. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36,** e105 (2008).

9. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–94 (1998).

10. Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* **11,** 473–83 (2010).

11. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18,** 440–5 (2002).

12. Trapnell, C. & Salzberg, S. How to map billions of short reads onto genomes. *Nat. Biotechnol.* **27,** 455–458 (2009).

13. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

15. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–11 (2009).

16. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–9 (2009).

17. Picard. at <http://picard.sourceforge.net/index.shtml>

18. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–40 (2010).

19. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

20. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–17 (2012).

21. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2013).

22. Li, J. & Tibshirani, R. Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* **22,** 519–36 (2013).

23. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14,** 91 (2013).

24. Reiter, L. T., Potocki, L., Chien, S., Gribskov, M. & Bier, E. A Systematic Analysis of Human Disease-Associated Gene Sequences In Drosophila melanogaster. 1114–1125 (2001). doi:10.1101/gr.169101.sophila

25. Lu, B. & Vogel, H. Drosophila models of neurodegenerative diseases. *Annu. Rev. Pathol.* **4,** 1184–1191 (2009).

26. Bilen, J. & Bonini, N. M. Drosophila as a model for human neurodegenerative disease. *Annu. Rev. Genet.* **39,** 153–171 (2005).

27. Rubin, L. L. & Haston, K. M. Stem cell biology and drug discovery. *BMC Biol.* **9,** 42 (2011).

28. Stadtfeld, M. & Hochedlinger, K. Induced pluripotency : history , mechanisms , and applications. 2239–2263 (2010). doi:10.1101/gad.1963910.Freely

29. Soldner, F. *et al.* Parkinson's Disease Patient-Derived Induced Pluripotent Stem Cells Free of Viral Reprogramming Factors. *Cell* **136,** 964–977 (2009).

30. Yang, Y. M. *et al.* Article A Small Molecule Screen in Stem-Cell-Derived Motor Neurons Identifies a Kinase Inhibitor as a Candidate Therapeutic for ALS. *Stem Cell* **12,** 713–726 (2013).

31. Fujita, K. The dark side of induced pluripotency Rethinking the sea-ice tipping point. doi:10.1029/1999GL006075

32. Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471,** 68–73 (2011).

33. Lefebvre, S. *et al.* Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* **80,** 155–65 (1995).

34. S, L. *et al.* Correlation between severity and SMN protein level in spinal muscular atrophy. *Nat. Genet.* **16,** 265–269 (1997).

35. Zhang, Z., Lotti, F., Dittmar, K., Younis, I. & Wan, L. SMN deficiency causes tissue-specific perturbations in the repertoire of snRNAs and widespread defects in splicing. *Cell* **133,** 585–600 (2008).

36. Bäumer, D. *et al.* Alternative splicing events are a late feature of pathology in a mouse model of spinal muscular atrophy. *PLoS Genet.* **5,** e1000773 (2009).

37. Lee, S. *et al.* Alternative Splicing Events Are a Late Feature of Pathology in a Mouse Model of Spinal Muscular Atrophy. **5,** (2009).

38. Chang, H. C.-H. *et al.* Modeling spinal muscular atrophy in Drosophila. *PLoS One* **3,** e3209 (2008).

39. Luo, L., Liao, Y. J., Jan, L. Y. & Jan, Y. N. Distinct morphogenetic functions of similar small GTPases: Drosophila Drac1 is involved in axonal outgrowth and myoblast fusion. *Genes Dev.* **8,** 1787–1802 (1994).

40. Dow, J. a T. Model organisms and molecular genetics for endocrinology. *Gen. Comp. Endocrinol.* **153,** 3–12 (2007).

41. Amaral, A. J. *et al.* Quality assessment and control of tissue specific RNA-seq libraries of Drosophila transgenic RNAi models. *Front. Genet.* **5,** 1–12 (2014).

42. Ebert, A., Yu, J., Rose, F. & Mattis, V. Induced pluripotent stem cells from a spinal muscular atrophy patient. *Nature* **457,** 277–280 (2008).

43. Anders, S. HTSeq: Counting reads in features with htseq-count. at <http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>

44. Anders, S. & Huber, W. Differential expression of RNA-Seq data at the gene level – the DESeq package. *Bioconductor Packag. Vignette* (2013).

45. R Development Core Team, R. R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput.* **1,** 409 (2011).

46. Falcon, S. & Gentleman, R. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23,** 257–8 (2007).

47. Carlson, M. org.Dm.eg.db: Genome wide annotation for Fly. R package version 2.14.0. at <http://www.bioconductor.org/packages/release/data/annotation/html/org.Dm.eg.db.html>

48. Carlson, M. org.Hs.eg.db: Genome wide annotation for Human. at <http://www.bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html>

49. Carlson, M. KEGG.db. at <http://www.bioconductor.org/packages/release/data/annotation/html/KEGG.db.html>

50. Carlson, M. GO.db: A set of annotation maps describing the entire Gene Ontology. at <http://www.bioconductor.org/packages/release/data/annotation/html/GO.db.html>

51. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21,** 3439–40 (2005).

52. Pollard, K., Dudoit, S. & Laan, M. Multiple Testing Procedures: the multtest Package and Applications to Genomics. *Stat. Biol. Heal.* 249–271 (2005). at <http://dx.doi.org/10.1007/0-387-29362-0_15>

53. Kasprzyk, A. BioMart: driving a paradigm change in biological data management. *Database (Oxford).* **2011,** bar049 (2011).

54. Gray, K. a *et al.* Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.* **41,** D545–52 (2013).

55. Marygold, S. J. *et al.* FlyBase: improvements to the bibliography. *Nucleic Acids Res.* **41,** D751–7 (2013).

56. Amit, M. *et al.* Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep.* **1,** 543–56 (2012).

57. Sueoka, N. On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. …* **1,** 582–592 (1962).

58. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12,** R18 (2011).

59. Park, J. W., Parisky, K., Celotto, A. M., Reenan, R. a & Graveley, B. R. Identification of alternative splicing regulators by RNA interference in Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 15974–9 (2004).

60. Neumüller, R. a *et al.* Genome-wide analysis of self-renewal in Drosophila neural stem cells by transgenic RNAi. *Cell Stem Cell* **8,** 580–93 (2011).

61. Wishart, T. M. *et al.* SMN deficiency disrupts brain development in a mouse model of severe spinal muscular atrophy. *Hum. Mol. Genet.* **19,** 4216–28 (2010).

62. Giavazzi, A., Setola, V., Simonati, A. & Battaglia, G. Neuronal-specific roles of the survival motor neuron protein: evidence from survival motor neuron expression patterns in the developing human central nervous system. *J. Neuropathol. Exp. Neurol.* **65,** 267–77 (2006).

63. Sanchez, G. *et al.* A novel function for the survival motoneuron protein as a translational regulator. *Hum. Mol. Genet.* **22,** 668–84 (2013).

64. Lotti, F. *et al.* An SMN-dependent U12 splicing event essential for motor circuit function. *Cell* **151,** 440–54 (2012).

65. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–78 (2012).

66. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–5 (2010).

67. Michael, D. & Manyuan, L. Intron—exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27,** 3219–3228 (1999).

68. Garcia, E. L., Lu, Z., Meers, M. P., Praveen, K. & Matera, a G. Developmental arrest of Drosophila survival motor neuron (Smn) mutants accounts for differences in expression of minor intron-containing genes. *RNA* **19,** 1510–6 (2013).

69. Lee, S. *et al.* Genome-wide expression analysis of a spinal muscular atrophy model: towards discovery of new drug targets. *PLoS One* **3,** e1404 (2008).

70. Comas, D., Petit, F. & Preat, T. Drosophila long-term memory formation involves regulation of cathepsin activity. **430,** 1–4 (2004).

71. Comas, D., Petit, F. & Preat, T. Drosophila long-term memory formation involves regulation of cathepsin activity. *Nature* **430,** 460–3 (2004).

72. Bridge, A. J., Pebernard, S., Ducraux, A., Nicoulaz, A.-L. & Iggo, R. Induction of an interferon response by RNAi vectors in mammalian cells. *Nat. Genet.* **34,** 263–4 (2003).

73. Zhang, Z., Lotti, F., Dittmar, K., Younis, I. & Wan, L. SMN deficiency causes tissue-specific perturbations in the repertoire of snRNAs and widespread defects in splicing. *Cell* **133,** 585–600 (2008).

74. Le, T. T. *et al.* SMNDelta7, the major product of the centromeric survival motor neuron (SMN2) gene, extends survival in mice with spinal muscular atrophy and associates with full-length SMN. *Hum. Mol. Genet.* **14,** 845–57 (2005).

75. Hughes, M. E., Grant, G. R., Paquin, C., Qian, J. & Nitabach, M. N. Deep sequencing the circadian and diurnal transcriptome of Drosophila brain. 1266–1281 (2012). doi:10.1101/gr.128876.111.1266

76. Daines, B. *et al.* The Drosophila melanogaster transcriptome by paired-end RNA sequencing. 315–324 (2011). doi:10.1101/gr.107854.110.21

77. Gan, Q., Chepelev, I., Wei, G., Tarayrah, L. & Cui, K. Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq. *Cell Res.* **20,** 763–783 (2010).

78. Xu, A. G. *et al.* Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq. *PLoS Comput. Biol.* **6,** e1000843 (2010).

6 – Appendixes

I – Alignment tables

a) *D. melanogaster* - Quantification of cDNA input, read quality filtering, mapping and protein coding annotation statistics of paired-end mRNA-seq libraries.

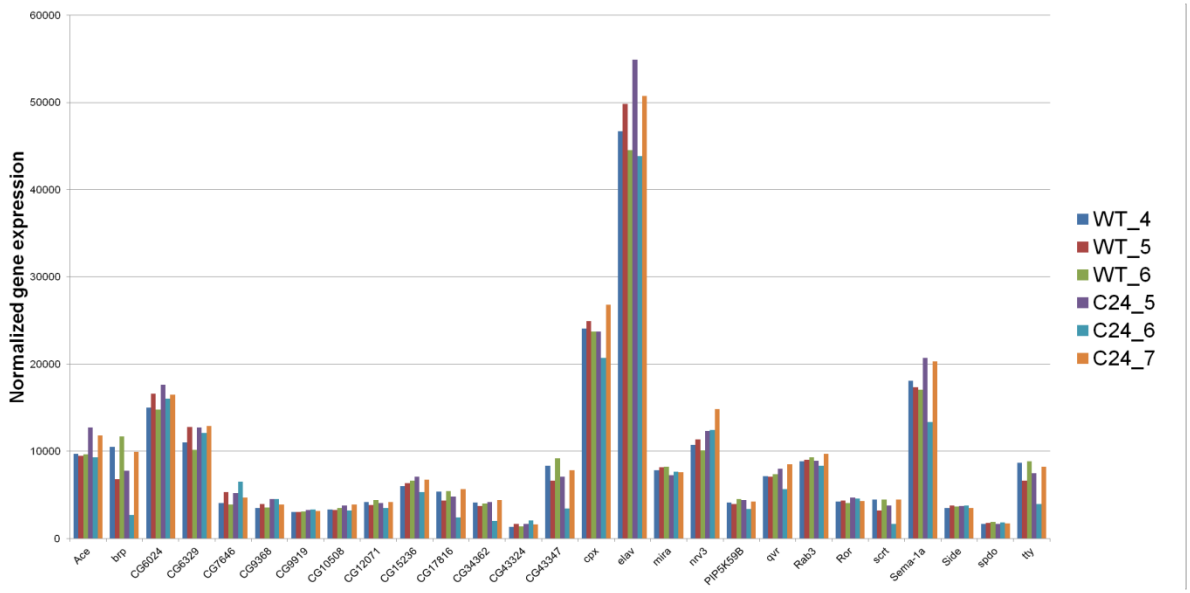
	cDNA used (ng)	Raw Reads (millions)	Uniquely aligned (millions)	Uniquely aligned w/out duplicates (millions)	Reads mapped to protein coding genes (millions)	Number of expressed genes	Normalized average read depth/gene
First Sequencing Batch							
C24 1	2	95,62	78,47	21,68 (27,63%)	15,77 (72,73%)	10725	2461
C24 2	6	112,11	87,95	53,45 (60,77%)	40,19 (75%)	11632	1353
C24 3	4	97,43	80,90	31,26 (38,64%)	21,85 (70%)	10975	1449
C24 4	3	111,77	85,32	51,07 (59,86%)	38,14 (75%)	12092	1322
WT 1	5	104,46	85,35	38,04 (44,57%)	29,42 (77%)	11985	1188
WT 2	2	106,03	87,69	23,52 (26,82%)	18,04 (76,7%)	10758	2186
WT 3	4	112,71	87,02	45,95 (52,80%)	33,83 (74%)	11712	1273
Second Sequencing Batch							
C24 5	13	86,25	65,83	48,39 (73,51%)	36,87 (76%)	11667	1506
C24 6	7	100,34	77,9	53,20 (68,29%)	40,47 (76%)	11669	1311
C24 7	12	98,35	77,40	53,17 (68,70%)	39,95 (75%)	11970	1471
WT 4	11	126,82	95,5	70,45 (73,77%)	53,95 (77%)	11964	1442
WT 5	10	84,38	66,12	47,21 (71,40%)	36,52 (77%)	11691	1426
WT 6	12	113,96	87,50	67,27 (76,88%)	53,40 (79%)	11984	1457
Third Sequencing Batch							
X7/C24 A	4	75,45	60,86	44,83 (73,66%)	33,60 (75%)	11665	1523
X7/C24 B	7	95,06	73,81	58,75 (79,60%)	43,81 (75%)	11935	1473
X7/C24 C	6	116,27	96,55	73,98 (76,62%)	56,70 (77%)	12096	1430
X7/C24 D	3	116,12	88,27	44,06 (49,92%)	33,03 (75%)	11386	1346

b) *H. sapiens* - Quantification of cDNA input, read quality filtering, mapping and protein coding annotation statistics of paired-end mRNA-seq libraries.

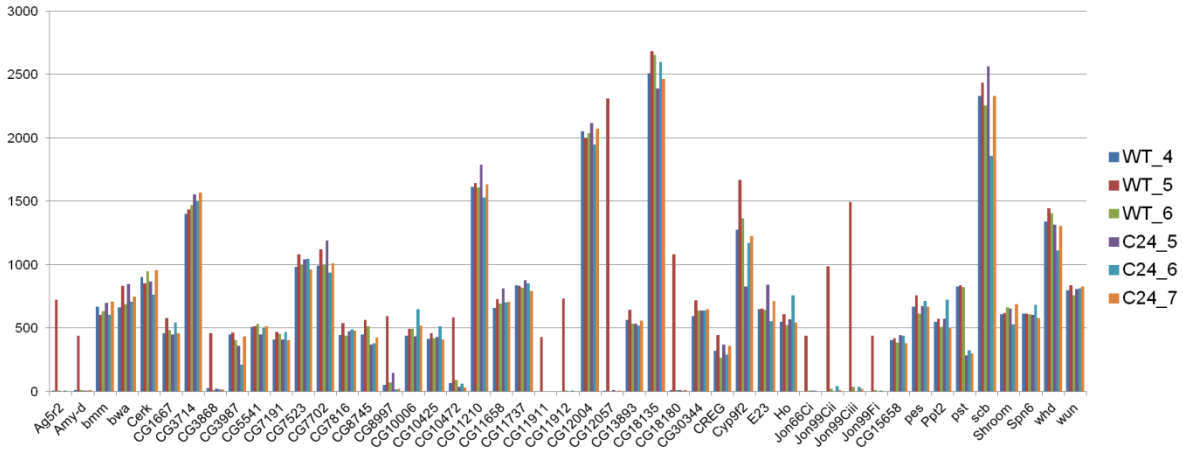
	Raw Reads (millions)	Uniquely aligned (millions)	Uniquely aligned w/out duplicates (millions)	Reads mapped to protein coding genes (millions)	Number of expressed genes	Normalized average read depth/gene
NS 1	116,42	101,05	71,26 (70,53%)	41,71 (59%)	19504	1338
NS 2	113,14	97,47	69,12 (70,93%)	41,03 (59%)	19316	1388
NS 3	112,2	97,54	69,82 (71,58%)	40,51 (58%)	19500	1340
shSMN2 1	80,86	70,11	50,5 (72,06%)	29,13 (58%)	19246	1311
shSMN2 2	99,46	86,48	63,16 (73,04%)	36,35 (58%)	19404	1353
shSMN2 3	88,6	77,27	55,06 (71,28%)	31,63 (57%)	19353	1344
SMAiPS 1	99,42	79,42	63,86 (80,41%)	36,45 (57%)	19669	1177
SMAiPS 2	87,76	70,29	54,94 (78,17%)	31,09 (57%)	19481	1192
SMAiPS 3	110,46	88,92	69,22 (88,92%)	40,91 (59%)	19749	1159

II – Tissue specific genes expression assessment for the WT_5 and C24_6 samples

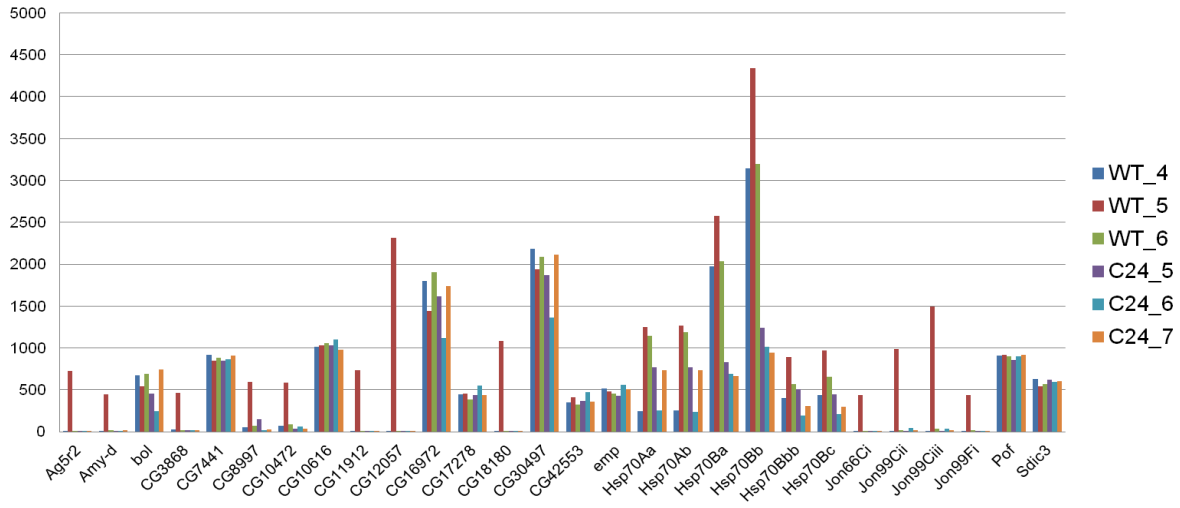
a) Neuronal exclusive genes with an expected high expression level



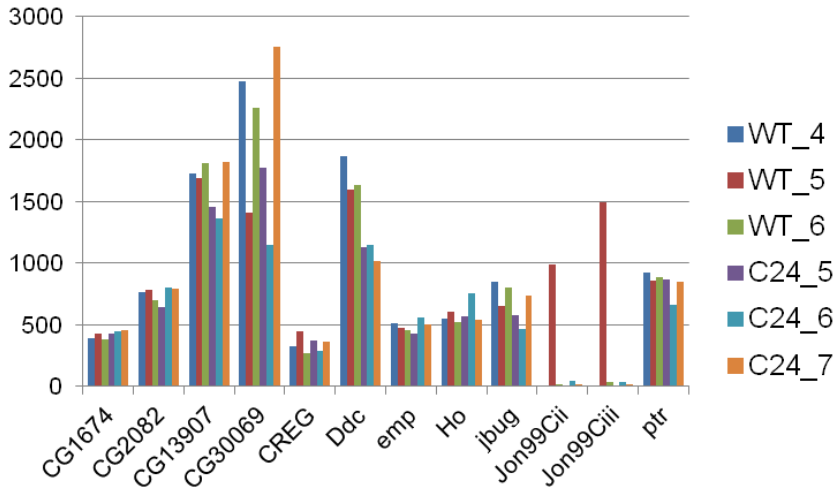
b) Digestive system genes with a significant expression (>400 reads) in WT_5 and C24_6



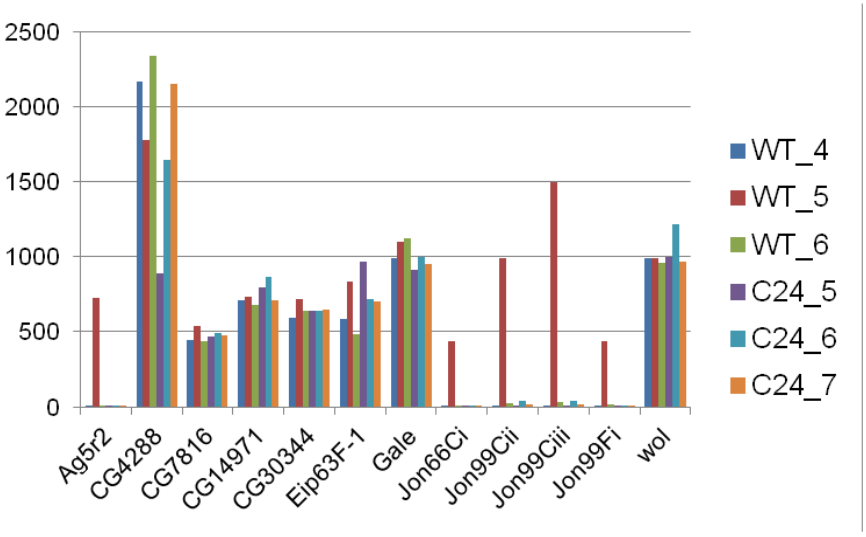
c) Imaginal Discs genes with a significant expression (>400 reads) in WT_5 and C24_6



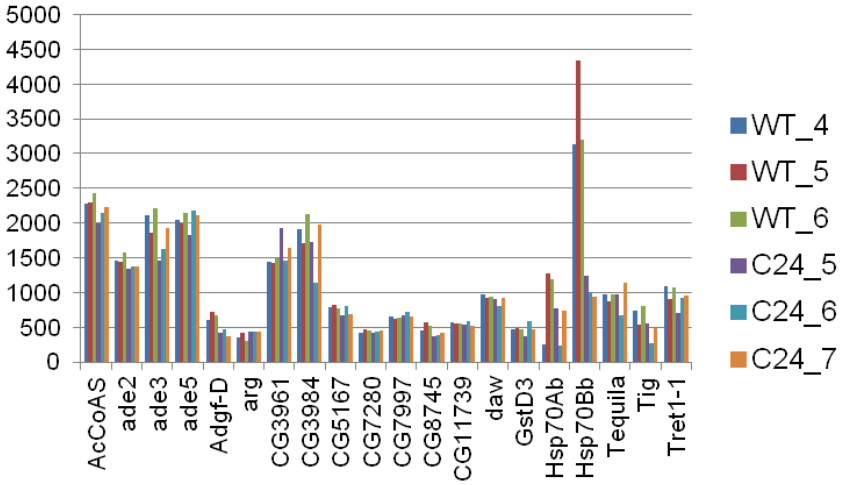
d) Carcass genes with a significant expression (>400 reads) in WT_5 and C24_6



e) Salivary gland genes with a significant expression (>400 reads) in WT_5 and C24_6



f) Fat body with a significant expression (>400 reads) in WT_5 and C24_6



III – GSEA results

D. melanogaster

a) Selected terms from the GSEA (BP) for the list of genes obtained in the WTxX7/C24 gene DEA.

GO ID	Gene Count	Term Size	Terms (down-regulated genes)	Ajusted p-value
GO:0006396	89	379	RNA processing	6,10E-13
GO:0022008	205	1257	neurogenesis	3,96E-11
GO:0042254	16	26	ribosome biogenesis	1,93E-09
GO:0006270	11	13	DNA replication initiation	6,06E-09
GO:0006418	19	40	tRNA aminoacylation for protein translation	1,65E-08
GO:0006399	27	76	tRNA metabolic process	2,25E-08
GO:0006364	15	29	rRNA processing	1,27E-07
GO:0006261	17	39	DNA-dependent DNA replication	3,05E-07
GO:0016071	60	300	mRNA metabolic process	1,50E-06
GO:0006281	24	80	DNA repair	2,77E-06
GO:0034660	17	50	ncRNA metabolic process	1,23E-05
GO:0000077	21	77	DNA damage checkpoint	5,78E-05
GO:0000398	44	226	mRNA splicing, via spliceosome	6,67E-05
GO:0000375	44	227	RNA splicing, via transesterification reactions	6,85E-05
GO:0006260	12	32	DNA replication	7,19E-05
GO:0071897	11	28	DNA biosynthetic process	9,73E-05
GO:0034504	19	70	protein localization to nucleus	0,000106218
GO:0006298	5	8	mismatch repair	0,000606523
GO:0006398	5	8	histone mRNA 3'-end processing	0,000606523

GO ID	Gene Count	Term Size	Terms (up-regulated genes)	Ajusted p-value
GO:0002165	150	576	instar larval or pupal development	2,62E-18
GO:0048569	122	451	post-embryonic organ development	2,88E-16
GO:0048699	157	730	generation of neurons	1,61E-11
GO:0048592	75	264	eye morphogenesis	4,46E-11
GO:0040011	103	485	locomotion	1,16E-07
GO:0031175	99	484	neuron projection development	9,93E-07
GO:0007399	236	1438	nervous system development	2,92E-06
GO:0048645	21	54	organ formation	3,49E-06
GO:0007409	61	280	axonogenesis	1,85E-05
GO:0000122	32	119	negative regulation of transcription from RNA polymerase II promoter	3,39E-05
GO:0007420	30	110	brain development	4,40E-05
GO:0007479	9	16	leg disc proximal/distal pattern formation	7,64E-05
GO:0048666	29	116	neuron development	0,000130668
GO:0007517	52	253	muscle organ development	0,00021038
GO:0045944	37	168	positive regulation of transcription from RNA polymerase II promoter	0,000380385
GO:0006376	5	7	mRNA splice site selection	0,000573674
GO:0051254	47	237	positive regulation of RNA metabolic process	0,000681019
GO:0051253	47	238	negative regulation of RNA metabolic process	0,000714817
GO:0006355	123	760	regulation of transcription, DNA-dependent	0,000742307
GO:0008045	14	45	motor neuron axon guidance	0,000758698

b) Selected terms from the GSEA (BP) for the list of genes obtained in the WT vs X7/C24 exon DEA

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0007411	106	187	axon guidance	1,90E-11
GO:0045664	37	56	regulation of neuron differentiation	5,95E-07
GO:0000375	110	227	RNA splicing, via transesterification reactions	6,20E-07
GO:0050768	23	29	negative regulation of neurogenesis	7,76E-07
GO:0022008	208	527	neurogenesis	6,00E-06
GO:0000398	99	212	mRNA splicing, via spliceosome	1,27E-05
GO:0016441	26	39	posttranscriptional gene silencing	2,25E-05
GO:0031647	23	34	regulation of protein stability	4,62E-05
GO:0043484	38	67	regulation of RNA splicing	5,40E-05
GO:0000381	33	56	regulation of alternative mRNA splicing, via spliceosome	6,08E-05
GO:0008582	22	33	regulation of synaptic growth at neuromuscular junction	8,47E-05
GO:0051961	18	25	negative regulation of nervous system development	9,08E-05
GO:0048675	19	27	axon extension	9,11E-05
GO:0045886	17	24	negative regulation of synaptic growth at neuromuscular junction	0,000171003
GO:0007399	22	38	nervous system development	0,000284095
GO:0048699	33	63	generation of neurons	0,000289724
GO:0000245	11	14	spliceosomal complex assembly	0,000542576
GO:0007416	15	22	synapse assembly	0,00056001
GO:0016246	12	16	RNA interference	0,000587362
GO:0030422	8	9	production of siRNA involved in RNA interference	0,000765454

H. sapiens

a) Selected terms from the GSEA (BP) for the list of genes obtained in the NS vs SMAiPS exon DEA

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0048667	87	558	cell morphogenesis involved in neuron differentiation	7,09E-06
GO:0048812	88	565	neuron projection morphogenesis	7,09E-06
GO:0016032	102	679	viral reproduction	7,09E-06
GO:0019048	57	359	virus-host interaction	0,00012509
GO:0044403	65	427	symbiosis, encompassing mutualism through parasitism	0,000135261
GO:0007411	56	354	axon guidance	0,000143802
GO:0030182	125	965	neuron differentiation	0,000173414
GO:0031333	17	68	negative regulation of protein complex assembly	0,000254384
GO:0031123	21	95	RNA 3'-end processing	0,000279915
GO:0006378	9	24	mRNA polyadenylation	0,000335514
GO:0016199	4	5	axon midline choice point recognition	0,000520432
GO:0008380	49	319	RNA splicing	0,000520432
GO:0022008	137	1115	neurogenesis	0,000534367
GO:0048011	36	220	nerve growth factor receptor signaling pathway	0,000745069
GO:0035385	3	3	Roundabout signaling pathway	0,000880441
GO:0022604	45	299	regulation of cell morphogenesis	0,000969473
GO:0010608	51	351	posttranscriptional regulation of gene expression	0,000983307

b) Selected terms from the GSEA (BP) for the list of genes obtained in the shSMN2 vs SMAiPS exon DEA

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0048667	87	558	cell morphogenesis involved in neuron differentiation	7,09E-06
GO:0048812	88	565	neuron projection morphogenesis	7,09E-06
GO:0016032	102	679	viral reproduction	7,09E-06
GO:0019048	57	359	virus-host interaction	0,00012509
GO:0007411	56	354	axon guidance	0,000143802
GO:0030182	125	965	neuron differentiation	0,000173414
GO:0031123	21	95	RNA 3'-end processing	0,000279915
GO:0042692	46	285	muscle cell differentiation	0,000298434
GO:0006378	9	24	mRNA polyadenylation	0,000335514
GO:0016199	4	5	axon midline choice point recognition	0,000520432
GO:0008380	49	319	RNA splicing	0,000520432
GO:0022008	137	1115	neurogenesis	0,000534367
GO:0048011	36	220	nerve growth factor receptor signaling pathway	0,000745069
GO:0035385	3	3	Roundabout signaling pathway	0,000880441
GO:0070271	97	762	protein complex biogenesis	0,000971005

c) GSEA (BP) for the list of genes obtained in the NS vs shsMN2 exon DEA

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0051704	148	1199	multi-organism process	1,69E-05
GO:0016032	91	669	viral reproduction	3,92E-05
GO:0044703	92	690	multi-organism reproductive process	5,93E-05
GO:0044403	60	412	symbiosis, encompassing mutualism through parasitism	0,000100112
GO:1901575	194	1739	organic substance catabolic process	0,000100112
GO:0019048	52	352	virus-host interaction	0,000159311
GO:0006397	55	390	mRNA processing	0,000250884
GO:0031175	86	682	neuron projection development	0,000250884
GO:0048002	28	173	antigen processing and presentation of peptide antigen	0,000743366
GO:0031123	18	95	RNA 3'-end processing	0,00096844
GO:0019884	26	160	antigen processing and presentation of exogenous antigen	0,00096844

d) GSEA (KEGG) for the list of genes obtained in the NS vs SMAiPS gene DEA

KEGGID	Gene Count	Term Size	Terms	Ajusted p-value
4510	132	197	Focal adhesion	1,90E-08
5200	199	320	Pathways in cancer	1,90E-08
4810	130	204	Regulation of actin cytoskeleton	1,35E-06
4512	60	83	ECM-receptor interaction	4,84E-06
4514	83	123	Cell adhesion molecules (CAMs)	4,84E-06
4360	86	129	Axon guidance	5,68E-06
4144	118	193	Endocytosis	4,35E-05
4670	72	110	Leukocyte transendothelial migration	7,95E-05
5414	59	87	Dilated cardiomyopathy	7,95E-05
5160	76	118	Hepatitis C	9,05E-05
5100	48	69	Bacterial invasion of epithelial cells	0,000132172
4380	70	109	Osteoclast differentiation	0,00017424
4020	101	167	Calcium signaling pathway	0,000180268
4530	79	126	Tight junction	0,000180268
5412	49	72	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0,000184997
5140	45	65	Leishmaniasis	0,000184997
4060	116	197	Cytokine-cytokine receptor interaction	0,000203461

e) GSEA (KEGG) for the list of genes obtained in the shSMN2 vs SMAiPS gene DEA

KEGGID	Gene Count	Term Size	Terms	Ajusted p-value
4510	128	193	Focal adhesion	6,14E-08
4360	90	128	Axon guidance	9,60E-08
4512	63	83	ECM-receptor interaction	9,63E-08
4514	83	120	Cell adhesion molecules (CAMs)	5,83E-07
5414	61	83	Dilated cardiomyopathy	7,47E-07
4810	126	199	Regulation of actin cytoskeleton	8,09E-07
5200	186	315	Pathways in cancer	1,72E-06
5410	53	76	Hypertrophic cardiomyopathy (HCM)	3,82E-05
5412	50	71	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	3,84E-05
4530	79	124	Tight junction	6,05E-05
5140	44	62	Leishmaniasis	7,31E-05
4670	68	107	Leukocyte transendothelial migration	0,000195687
5160	73	118	Hepatitis C	0,000356237
5131	41	60	Shigellosis	0,000388376
4971	46	69	Gastric acid secretion	0,000388376
5146	58	91	Amoebiasis	0,000401767
4020	99	170	Calcium signaling pathway	0,000603004

Human-Mouse orthologs

a) Selected terms (lowest adj-p) from the GSEA (BP) for the list of human genes obtained in the overlap between NS vs SMAiPS and Zhang et al.

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0030199	10	37	collagen fibril organization	1,56E-10
GO:0043062	14	200	extracellular structure organization	5,54E-07
GO:0048667	20	558	cell morphogenesis involved in neuron differentiation	2,13E-05
GO:0048812	20	565	neuron projection morphogenesis	2,13E-05
GO:0022610	26	905	biological adhesion	2,13E-05
GO:0007411	15	354	axon guidance	3,71E-05
GO:0048646	35	1546	anatomical structure formation involved in morphogenesis	4,86E-05
GO:0030030	25	921	cell projection organization	5,83E-05
GO:0030155	12	259	regulation of cell adhesion	9,72E-05
GO:0030182	25	965	neuron differentiation	0,0001034
GO:0071230	5	35	cellular response to amino acid stimulus	0,000113901
GO:0001101	7	85	response to acid	0,000113901
GO:0032990	20	683	cell part morphogenesis	0,000113901
GO:0042330	17	559	taxis	0,000242821
GO:0009611	24	983	response to wounding	0,000242821
GO:0022008	26	1115	neurogenesis	0,000242821
GO:0071840	65	4129	cellular component organization or biogenesis	0,000242821
GO:0043206	3	10	extracellular fibril organization	0,00031966
GO:0007160	8	143	cell-matrix adhesion	0,00031966
GO:0050896	85	6008	response to stimulus	0,00031966

b) Selected terms (lowest adj-p) from the GSEA (BP) for the list of human genes obtained in the overlap between NS vs SMAiPS and Bäumer et al.

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0009628	21	727	response to abiotic stimulus	0,000290805
GO:0097411	2	2	hypoxia-inducible factor-1alpha signaling pathway	0,000497182
GO:0006817	3	10	phosphate ion transport	0,000497182
GO:0046479	3	10	glycosphingolipid catabolic process	0,000497182
GO:0044255	20	780	cellular lipid metabolic process	0,000497182
GO:0019216	9	197	regulation of lipid metabolic process	0,000564819
GO:0007173	8	158	epidermal growth factor receptor signaling pathway	0,000564819
GO:2000377	5	56	regulation of reactive oxygen species metabolic process	0,000564819
GO:0009409	4	31	response to cold	0,000564819
GO:0048511	8	169	rhythmic process	0,000608745
GO:0042136	3	14	neurotransmitter biosynthetic process	0,000608745
GO:0045017	9	219	glycerolipid biosynthetic process	0,00063572
GO:0048011	9	220	nerve growth factor receptor signaling pathway	0,00063572
GO:0009725	16	610	response to hormone stimulus	0,000684042
GO:0008543	7	138	fibroblast growth factor receptor signaling pathway	0,000684042
GO:0046514	3	17	ceramide catabolic process	0,00076443
GO:0019318	9	236	hexose metabolic process	0,000806154
GO:1901652	11	341	response to peptide	0,000806602
GO:0007163	6	111	establishment or maintenance of cell polarity	0,000972661

Human-*Drosophila* orthologs

a) Selected terms (lowest adj-p) from the GSEA (BP) for the human gene list obtained from the overlap of NSxSMAiPS DE genes with the X7/C24 DE genes.

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0048699	127	1048	generation of neurons	3,68E-05
GO:0048489	17	66	synaptic vesicle transport	0,000158548
GO:0030204	15	55	chondroitin sulfate metabolic process	0,000198665
GO:0045595	115	983	regulation of cell differentiation	0,00021977
GO:0030155	40	259	regulation of cell adhesion	0,000286032
GO:0005975	91	746	carbohydrate metabolic process	0,000286032
GO:0051960	59	441	regulation of nervous system development	0,00043257
GO:0048731	96	861	system development	0,000451936
GO:0015031	134	1212	protein transport	0,000451936
GO:0048646	165	1546	anatomical structure formation involved in morphogenesis	0,000451936
GO:0071372	5	8	cellular response to follicle-stimulating hormone stimulus	0,000462707
GO:0006096	14	58	glycolysis	0,000462707
GO:0042593	22	117	glucose homeostasis	0,000462707
GO:0030029	59	451	actin filament-based process	0,000462707
GO:0006461	90	760	protein complex assembly	0,000462707
GO:0016192	70	570	vesicle-mediated transport	0,00046332
GO:0031109	12	47	microtubule polymerization or depolymerization	0,00057094
GO:0007417	81	677	central nervous system development	0,00057094
GO:0031175	82	689	neuron projection development	0,00057094
GO:0006082	107	947	organic acid metabolic process	0,00057094
GO:0019320	17	85	hexose catabolic process	0,00069096
GO:0045766	17	85	positive regulation of angiogenesis	0,00069096
GO:0042254	25	149	ribosome biogenesis	0,00069096
GO:0045664	44	321	regulation of neuron differentiation	0,00069096

b) Selected terms (lowest adj-p) from the GSEA (BP) for the fly gene list obtained from the overlap of NSxSMAiPS DE genes with the X7/C24 DE genes.

GO ID	Gene Count	Term Size	Terms	Ajusted p-value
GO:0048569	78	451	post-embryonic organ development	4,47E-08
GO:0048812	79	483	neuron projection morphogenesis	3,12E-07
GO:0007269	29	112	neurotransmitter secretion	6,09E-07
GO:0046903	39	179	secretion	6,09E-07
GO:0040011	78	485	locomotion	6,09E-07
GO:0007268	52	276	synaptic transmission	6,13E-07
GO:0007389	52	281	pattern specification process	6,13E-07
GO:0003001	30	121	generation of a signal involved in cell-cell signaling	8,50E-07
GO:0016192	72	443	vesicle-mediated transport	9,76E-07
GO:0035637	52	283	multicellular organismal signaling	1,19E-06
GO:0042067	19	59	establishment of ommatidial planar polarity	2,15E-06
GO:0001738	27	108	morphogenesis of a polarized epithelium	2,44E-06
GO:0051674	47	251	localization of cell	2,44E-06
GO:0001709	32	142	cell fate determination	2,60E-06
GO:0016477	44	230	cell migration	2,66E-06
GO:0048729	54	308	tissue morphogenesis	2,66E-06
GO:0007164	25	97	establishment of tissue polarity	3,03E-06
GO:0048737	61	372	imaginal disc-derived appendage development	4,73E-06
GO:0006928	62	381	cellular component movement	4,86E-06

IV – Protocols

Protocol 1 - Data Filtering

1) The protocol starts out with 3 files for each read of each condition's replicates: condition replicate read ID.fastq, condition replicate read SEQ.fastq and condition replicate read QS.fastq which contain, respectively, the ID from the read, the read's sequence and the quality score of each nucleotide. These are joined into one table with a simple bash command:

$ paste <ID.fastq> <SEQ.fastq> <QS.fastq> > TABLE.fastq

2) These tables are then filtered with a perl script created for this purpose, which removes the homopolymers equal or longer than 50% of the read's total size, non called bases and reads with a QS lower than 30:

$ perl filter_RNAseq_Harvard2.pl <input_TABLE.fastq> <output_TABLE_filtered.fastq>

3) From these filtered tables, we now need to know which reads still have their pair, which can be done with a script also created for this purpose. First we need to extract the ID's from each read and join them in pairs, for each replicate.

$ cut -f <Read_1_TABLE_filtered.fastq> > IDlist_Read1

$ cut -f <Read_2_TABLE_filtered.fastq> > IDlist_Read2

$ cat <IDlist_Read1 IDlist_Read2> >> IDlist_both_reads

4) Finally, a Perl in-house script identifies and removes reads which don't have both pairs featured after filtering:

$ perl prepare_fastq_RNAseq_paired_ends_line.pl <IDlist_both_reads> <Read1_TABLE_filtered.fastq> <Read1_filtered_TABLE_paired.fastq> 1:N:0:

$ perl prepare_fastq_RNAseq_paired_ends_line.pl <IDlist_both_reads> <Read2_TABLE_filtered.fastq> <Read2_filtered_TABLE_paired.fastq> 2:N:0:

Protocol 2 - BWA alignment, gene count with HTseq and DEA/visualization with DESeq/DEXSeq

1) Retrieve the reference genome in fasta format, in order to create the BWT indexes:

$ bwa index [-a is/bwswt] <genome.fa>

2) The aln command finds the suffix array coordinates of the input reads. Using read 1 as an example:

$ bwa aln [-n edit distance] [-t nThreads] <index> <Read1_filtered_TABLE_paired.fastq> <out.sai>

3) The sampe command converts SA coordinates to chromosomal coordinates, generating the alignments in SAM format, given paired-end reads.

$ bwa sampe [-n maxHitPaired] [-N maxHitDistance] <index> <Read1.sai> <Read2.sai> <Read1_filtered_TABLE_paired.fastq> <Read2_ filtered_TABLE_paired.fastq> > <paired.sam>

4) Removal of PCR duplicates is made with samtools. Since the software which removes the duplicates needs to work with .bam files, first we need to convert the file from .sam to its binary form (.bam), sort it by chromosome, run the duplicate removal tool, sort it by read pairs and finally, convert it back to .sam

$ samtools view [-bS .bam from .sam] [-o output] paired.bam paired.sam &

$ samtools sort paired.bam paired_sorted &

$ samtools rmdup paired_sorted.bam paired_no_duplicates.bam &

$ samtools sort [-n by read] paired_no_duplicates.bam paired_clean &

$ samtools view [-h bam to sam] [-o output] paired_clean.sam paired_clean.bam &

5) HTSeq is then used to do the gene count, using one of two approaches: union and intersect_strict. Note that the default option for –m (mode) is union.

$ htseq-count [-a skip QS < than n] [-s strand specific? (yes/no)] [-m union/intersection-strict] <paired.sam> <genome.gff> > <output.table>

6) DESeq/DEXSeq analysis. An R script for each package was created based on the bioconductor vignettes, which normalizes library sizes and analyses the differential gene/isoform expression between samples and allows for graphical visualization of the data.

Protocol 3 – *GAL4* quantification

1) Retrieve the *GAL4* sequence (Gene ID 855828; accession NC_001148.4) in fasta format and create a BWT index for alignment

   $ bwa index [-a is/bwswt] <genome.fa>

2) Using the files created in Protocol 1, follow Protocol 2's steps 1) through 4) using the newly created reference "genome"

3) Count the aligned to *GAL4* reads using the command line and write them to a text file

   $ awk '$3=="gi|330443753:c82356-79711" { print $0 }' GAL4search_EXAMPLE.sam > GAL4search_EXAMPLE_aligned &